Sparse and structured decomposition of audio signals on hybrid dictionaries using musical priors

Hélène Papadopoulos, Matthieu Kowalski

Sparse and structured decomposition of audio signals on hybrid dictionaries using musical priors

Hélène Papadopoulos[a)] and Matthieu Kowalski
Laboratoire des Signaux et Systèmes
UMR 8506, CNRS-SUPELEC-Univ Paris-Sud
91172 Gif-sur-Yvette Cedex, France

(Dated: September 16, 2013)

This paper investigates the use of musical priors for sparse expansion of audio signals of music, on an overcomplete dual-resolution dictionary taken from the union of two orthonormal bases that can describe both transient and tonal components of a music audio signal. More specifically, chord and metrical structure information are used to build a structured model that takes into account dependencies between coefficients of the decomposition, both for the tonal and for the transient layer. The denoising task application is used to provide a proof of concept of the proposed musical priors. Several configurations of the model are analyzed. Evaluation on monophonic and complex polyphonic excerpts of real music signals shows that the proposed approach provides results whose quality measured by the signal-to-noise ratio is competitive with state-of-the-art approaches, and more coherent with the semantic content of the signal. A detailed analysis of the model in terms of sparsity and in terms of interpretability of the representation is also provided, and shows that the model is capable of giving a relevant and legible representation of Western tonal music audio signals.

PACS numbers: 43.75.Xz, 43.75.Zz, 43.60.Hj, 43.60.Pt

I. INTRODUCTION

We describe in this paper a novel approach for *structured sparse* decomposition of a music signal in an overcomplete time-frequency hybrid dictionary. Within a Bayesian framework, we propose to incorporate musical priors in order to built signal representations that take into account some "structural" information and that are more suitable to music than existing methods that are based on physical signal properties. For this, we take advantage of the recent work that have been done on chord estimation and beat tracking in the context of music content indexing. The model we propose is inspired from previously proposed Bayesian models for time-frequency inverse modeling of non-stationary signals (Wolfe *et al.*, 2004) or sparse linear regression in unions of bases (Févotte *et al.*, 2008), but presents an essential difference in the way dependencies between coefficients of the representation are modeled, using the newly introduced musical priors. One of the goal of this work is to show that ideas that have emerged in two related but distinct communities, *Sparsity* and *Music Information Retrieval*, can be exploited jointly to open new perspective for audio signal processing. Sparse representations have been used as a basis for extracting high-level information for MIR applications, such as note extraction in (Davies and Daudet, 2006), or beat tracking, chord recognition and musical genre classification in (Ravelli *et al.*, 2010). Our approach is somewhat different: we propose to incorporate music content information in order to build structured sparse representations that are tailored to the analyzed music signal and legible from a musical point of view.

A. Structured sparse representation

The problem of representing an audio signal using a time-frequency dictionary has been given a lot of attention these last few years. The specific problem we consider here is finding an approximation of a music audio signal as a linear combination of elementary waveforms (also called *atoms*) of a suitably chosen dictionary.

Musical signals are intrinsically structured. A particularity of musical signals is that, very often, several types of components are superimposed such as, among other features, tonal components (the partials of the notes, that are characterized by sinusoids with slowly varying amplitude and frequency) and transients (the attacks of the notes, that correspond to events well-localized in time). This is illustrated in Figure 1 that represents the spectrogram of a glockenspiel signal. The tonals appear as thin horizontal lines whereas the transients appear as sharp vertical lines. These various components may have significantly different behaviors in terms of time-frequency localization. For instance, fast varying transients require short analysis window length, whereas low varying tonals require long windows. Thus, they cannot be optimally represented within the same basis. The Balian-Low theorem (Low, 1985) states that redundancy is necessary for having well-localized functions both in time and frequency for transforms based on local Fourier analysis. This is why *hybrid* models, allowing a simultaneous representation of different components have been proposed (Hamdy *et al.*, 1996; Verma and Meng, 2000; Daudet and Torrésani, 2002; Molla and Torésani, 2005). These models consider redundant (or overcomplete) dictionaries that are constructed by the concatenation of several families or bases (usually time-frequency atoms, such as Gabor atoms or local cosines; or time-scale atoms, such

[a)] Author to whom correspondence should be addressed. Electronic mail: helene.papadopoulos@lss.supelec.fr.

as wavelets). Typically, the resulting dictionaries contain elements with various time-frequency characteristics and are more flexible than orthonormal bases (Chen *et al.*, 1998).

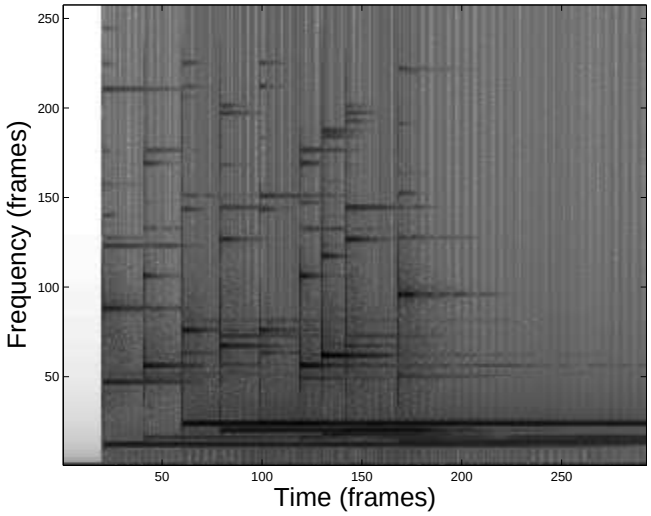

FIG. 1. Time-frequency representation of a recording of a glockenspiel excerpt. The vertical lines correspond to the attacks of the notes and the horizontal lines correspond to the partials.

Among the various existing transforms, lapped transforms such as the Modified Discrete Cosine Transform (MDCT) (Mallat, 1998; Malvar, 1990) is a standard choice for the bases (Févotte *et al.*, 2006; Kowalski and Torrésani, 2008). This transform is very popular, in particular in high quality audio coding and signal compression applications, because it allows an orthogonal time-frequency transform without blocking effects. Following these approaches, we consider in this work a dictionary built as the union of two MDCT bases with different time-frequency resolutions. The narrow band basis – with long time resolution – is used to estimate the tonal parts of the signal, and the wide band basis – with short time resolution – is used to estimate the transient parts. Such a dictionary is overcomplete since the number of elements of the dictionary is greater than the length of the signal. The expansion of the signal with respect to the dictionary is thus not unique. *Sparsity* may be used as a selection criterion for finding the expansion coefficients, in the sense that only a few coefficients of the decomposition of the signal on the bases are significantly nonzero. The signal can thus be well approximated by a limited number of coefficients. This problem is often referred to as *sparse regression*. Sparsity has become a fundamental concept in diverse areas of modern signal processing. It is, for instance, an essential ingredient of popular coding standards such as the MPEG-1 layer III ("MP3") or the MPEG-2 AAC. A review of sparse representations for musical signals and their applications can be found in (Daudet and Torrésani, 2006; Plumbley *et al.*, 2010).

A common approach to find a sparse expansion of signals in overcomplete dictionaries consists of minimizing the ℓ_1 norm of the expansion, and is known as *basis pursuit* (Chen *et al.*, 1998), or LASSO (Tibshirani, 1996). Other methods include variational approaches (Kowalski, 2009), probabilistic approaches (Kowalski and Torrésani, 2008), greedy methods, such as the Matching Pursuit algorithm (Mallat and Zhang, 1993; Daudet, 2010) and its variants (the Orthogonal Matching Pursuit (Mallat, 1998; Pati *et al.*, 1993), the Molecular Matching Pursuit (Daudet, 2006b)), or Bayesian formulations as for instance EM-based algorithms (Figueiredo, 2003). In the framework of Bayesian variable selection, MCMC (Markov chain Monte Carlo) type approaches have been proposed (Févotte and Godsill, 2006; Févotte *et al.*, 2008). One of the main advantages of the MCMC techniques is their robustness because they scan the whole of the posterior distribution and thus are unlikely to fall into local minima. However, this is done at the expanse of high computational cost.

The concept of *structured approximation* has been introduced from the observation that significant coefficients are not isolated but tend to form "clusters" in the index space. As mentioned above, audio music signals are well-structured. In the time-frequency plane, the partials of the notes will generate horizontal lines localized in frequency, whereas the attacks of the notes and the percussive sounds will generate vertical lines localized in time. Ideally, this structure should be reflected in the signal decomposition, so that the coefficients have physical interpretability and are more meaningful than isolated coefficients from an analysis perspective. Interpretability is a key concept in sparsity[1]. A signal representation where coefficients can be explained from a theoretical or a physical point of view can help assessing the model accuracy, and provides a suitable representation for higher-level tasks. For instance in music signal analysis, a time-frequency representation where coefficients can be physically interpretable as being part of the tonal layer may be very useful to the multi-f0 estimation task (Yeh *et al.*, 2010). This is why we are interested in finding a signal approximation that is not only sparse, but also *structured*, by considering dependencies between significant coefficients. Previous approaches that use unstructured priors, such as Bernoulli models have shown that they generate isolated coefficients with high amplitude in both bases (Févotte and Godsill, 2006; Kowalski and Torrésani, 2008). These components do not have any physical or musical meaning and are usually perceived as "musical artifacts" or "musical noise" in the reconstructed signal. Considering dependencies between atoms coefficients and using structured priors allows reducing the number of such undesirable components. Various approaches have been proposed for introducing dependencies between coefficients in the time-frequency domain. These approaches aim at exploiting the fact that significant coefficients tend to be organized into clusters, which results from persistence properties of time-frequency representations. Structures can be modeled directly in the coefficients themselves, such as in (Kowalski, 2009). However, dependencies are often introduced in the time-frequency indices, rather than directly in the coefficients. This results into hierarchical models in which both the coefficients and the addresses of the significant

coefficients have to be modeled.

Among existing approaches, physical properties resulting in persistency over frequency of the transient layer can be modeled using structured hierarchical Bernoulli models on a dictionary built as the union of two MDCT bases with different time-frequency resolutions (Kowalski and Torrésani, 2008), binary Markov trees (Crouse *et al.*, 1998; Molla and Torésani, 2005), or dyadic trees of wavelet coefficients used with wavelet bases (Daudet and Torrésani, 2002); persistency of the tonal layer can be favored using Markov chains, as proposed in (Molla and Torésani, 2005) in the case of a MDCT base; in (Févotte *et al.*, 2008), structural constraints on the coefficients that rely on physical properties of the signal are imposed for both layers. Persistencies of time-frequency coefficients of musical signal are modeled using two types of Markov chains. It results in a "horizontal structure" for the tonal layer and a "vertical structure" for the transient layer.

Enforcing structure between expansion coefficients can be managed using sequential approaches (Daudet and Torrésani, 2002; Daudet, 2004; Molla and Torésani, 2005) that first identify the tonal layer using the first basis, and then estimate the transient components from the residual, using the second basis. In (Daudet and Torrésani, 2002; Daudet, 2004), tonal and transient components are expanded sequentially into local cosine and dyadic wavelets bases respectively. The method does not rely on any prior segmentation of the signal. For each layer, only the largest coefficients in each time frame are retained based on threshold values that are estimated adaptively in a quantization stage. In the framework of audio coding, (Molla and Torésani, 2005) describes an hybrid model for the expansion of audio signals considering a redundant dictionary made out of the union of local cosine and wavelet bases. A recursive scheme is proposed to estimate the two layers, that relies on the assumption that the cardinalities of the significance maps have to be known. A priori estimates for the relative sizes of the tonal and transient layers are obtained based on an algorithm that determines local transientness of audio signals (Molla and Torésani, 2004). The approach is used to develop an hybrid audio coder that does not rely on prior (time) segmentation of the signal.

As stressed in (Daudet, 2006a), sequential approaches suffer from two limitations. First, errors in a step are systematically propagated into the next estimation stage and thus bias the estimates of the other components. Second, the choice of a threshold that allows discriminating large significant from small residual coefficients is difficult. An alternative to sequential approaches is the simultaneous approach of both layers, as proposed in (Févotte *et al.*, 2008).

Starting from this approach, we build a structured model for sparse signal decomposition within a Bayesian framework. The originality of our work is that we model dependencies between the expansion coefficients by using priors that are based on musical information.

B. Content-based music information retrieval

Up to now, additional structure constraints that have been added rely on physical properties of the signal. The recent advances in automatic extraction of content information from audio music signals in the field of Music Information Retrieval (MIR) offers an interesting alternative. Content-based music information retrieval deals with the extraction and processing of meaningful information from musical audio. Techniques developed for searching, retrieving, organizing and interacting in a personalized way with large databases of music signals are often based on the use of musical descriptors that are extracted from the signal, such as the key, the chord progression, the melody or the instrumentation. Musical content information can be used to build structured priors that reflect the content of the signal. For instance, as we propose in this paper, the chord progression provides information about the notes that are present in the signal and can be used to build a prior for the tonal layer. Similarly, the position of the beats is related to the transients and can be used to build a prior for the transient layer. These concepts are introduced in what follows.

1. Chord estimation

The chord progression of a piece of music is a very important descriptor because it characterizes its harmonic structure. Here, we want to work directly on audio. The symbolic transcription (the score) of a piece of music is not always available, especially in music genres such as jazz music where there is a large part devoted to improvisation. In addition, algorithms that extract a transcription from an audio signal, such as multi-f0 estimation algorithms (Yeh *et al.*, 2010), are still limited and costly. However, numbers of recent work have shown that it is possible to accurately extract a robust representation of the harmonic content without the use of transcription algorithms. Estimating the chord progression of an audio signal has thus become a very popular task in MIR (Sheh and Ellis, 2003; Bello and Pickens, 2005; Harte and Sandler, 2005; Papadopoulos and Peeters, 2011).

The output of a chord estimation algorithm consists in a progression of chords chosen among a given chord lexicon, that is very often limited to the 24 major and minor triads. Chord estimation on real signals has been favored by the use of the *chroma* features (Wakefield, 1999) or Pitch Class Profiles (Fujishima, 1999), which are traditionally 12-dimensional vectors, with each dimension corresponding to the intensity associated with one of the 12 semitone pitch classes (chroma) of the Western tonal music scale, regardless of octave. The temporal sequence of chroma vectors over time is known as *chromagram*. Conceptually, the chromagram is a frequency spectrum folded into a single octave. Pooling the spectrum into twelve bins that correspond to the twelve pitch classes of the equal-tempered scale results in a signal representation that allows identifying pitches by an octave. Each chord may be characterized by the semitone pitch classes

or chroma that correspond to the notes it is composed of. The use of such a mid-level representation overcomes the problem of automatic transcription.

Various approaches for chroma computation exist. Although they present some variances in the implementation, they follow in general the same guideline that consists of two main steps:

1. First, a semitone pitch class spectrum (SPS), that is a log-frequency representation of the spectral content of the music audio signal, is constructed. It is expressed in a MIDI-note scale and is in general either computed from the Fourier transform or from the constant-Q transform (Brown, 1991).

2. Secondly, the semitone pitch spectrum is mapped to the chroma vectors. For this, the semitones in octave distance are added up to pitch classes.

The chromagram computation may include some other steps such as a pre-processing step that separates harmonic and noise components, a filtering step that smoothes the chromagram or a post-processing normalization step that makes the chromagram invariant to dynamics.

We rely in this paper on a chromagram computation method, described in (Papadopoulos and Peeters, 2011), that is based on a constant-Q transform applied on a downsampled signal.

For chord estimation, we rely on the model proposed in (Papadopoulos and Peeters, 2007, 2011), that is based on chord templates and hidden Markov models. We briefly described here the concepts that are used in the rest of the paper. The front-end of our model is based on the extraction of a chromagram that represents the audio signal. The chord progression is then modeled as an ergodic 24-states HMM, each hidden state corresponding a chord of a the chord lexicon (CM, ..., BM, Cm, ..., Bm), and the observations being the chroma vectors. The observation chord symbol probabilities are obtained by computing the correlation between the observation vectors (the chroma vectors) and a set of chord templates which are the theoretical chroma vectors corresponding to the $I = 24$ major and minor triads. A state-transition matrix based on musical knowledge (Noland and M., 2006) is used to model the rules from which the transitions between chords result. The chord progression over time is estimated in a maximum likelihood sense by decoding the underlying sequence of hidden chords $S = (s_1, s_2, \ldots, s_N)$ from the sequence of observed chroma vectors using the Viterbi decoding algorithm.

2. Beat tracking

Beat tracking is a challenging problem that has been addressed in a large number of works because beat information is used in many applications in music signal processing, such as music analysis, score alignment or cover version identification. Numerous good overviews on the problem of beat tracking are available, such as for instance (Dixon, 2007; Scheirer, 1998; Klapuri *et al.*, 2006; Davies and Plumbley, 2007). In the present work, we use the beat tracker proposed in (Peeters and Papadopoulos, 2011) as a front end of the system. Briefly, this approach aims at simultaneously estimating beat locations with downbeat locations from an audio file. A probabilistic framework in which the time of the beats and their associated beat-positions-inside-a-measure role, hence the downbeats, are considered as hidden states and are estimated simultaneously using signal observations. For this, a reverse Viterbi algorithm that decodes hidden states over beat-numbers is proposed. A beat-template is used to derive the beat observation probabilities. A machine-learning method, the Linear Discriminant Analysis, allows estimating the most discriminative beat-templates. Two kinds of observations are proposed to derive the beat-position-inside-a-measure observation probability: the variation over time of chroma vectors and the spectral balance. This methods was ranked first for the McKinney Collection test-set during the MIREX 2009 beat tracking contest. We refer the reader to (Peeters and Papadopoulos, 2011) for more details.

C. Contributions

Sparse representations of signals have recently proved to be useful for a wide range of applications in signal processing, such as denoising (Févotte *et al.*, 2006), coding and compression (Daudet *et al.*, 2004; Ravelli *et al.*, 2008), source separation (Benaroya *et al.*, 2006; Févotte and Godsill, 2006) or music transcription (Blumensath and Davies, 2004). Here, we focus on the task of denoising an excerpt of musical audio. The approach we propose is in many respects related to previously proposed MCMC schemes for nonlinear approximation in hybrid dictionaries of waveforms. However, a main difference is that we aim at providing a multilayered signal decomposition that fits the music signal, in which the layers can well explain the signal and reflect its music content and can provide more relevant semantic information. By incorporating musical priors, we built a model that is particularly well adapted to music and fits the intrinsic nature of Western tonal music. The denosing application is used as a proof of concept for the description of new musical priors introduced in the paper, and we thus focus on assessing the relevance of the new priors rather than demonstrating the superiority of the method in terms of denoising results (although it is competitive with respect to signal to noise ratio to the state-of-the-art). In this context, we systematically compare our model, in which structural constraints on the coefficients are based on musical prior, to the model proposed (Févotte *et al.*, 2008), and in which structural constraints on the coefficients rely on physical properties of the signal are imposed for both layers, reaching the state-of-the-art in terms of SNR results.

Preliminary results of the proposed approach can be found in (Papadopoulos and Kowalski, 2011). In this article, we propose significant improvements to the signal

model, in particular by presenting a structured model for the transient layer; we include the result of new experiments; finally we present a detailed analysis of the model and case-study examples.

The remainder of this paper is structured as follows. In Section II, we present our model for sparse signal decomposition on hybrid dictionaries that incorporates musical priors; our main contribution is described in part II.C where we specify our formulation of prior dependence structures in the time-frequency plane. We briefly address the problem of parameters estimation in Section III. In Sections IV and V, we present and discuss the simulation results of our model. Conclusions and perspectives for future work are given in Section VI.

II. SIGNAL MODEL

This section introduces first the mathematical model used to represent the audio signal, and then defines the priors chosen in a Bayesian context. Particularly, the new musical priors based on the *chromagram* and the beat locations are exposed in Section II.C.

A. Model

In this part, we describe our model for signal decomposition with sparse constraint on a *hybrid* dictionary of elementary waveforms (Daudet and Torrésani, 2002). The dictionary is constructed as the union of two orthonormal bases with different time-frequency resolution that account respectively for the tonal and the transient parts of the signal. We consider a tree-layer signal model of the form:

$$signal = tonals + transients + residual \ .$$

Let $V = \{v_n, n = 1, \ldots, N\}$ and $U = \{u_m, m = 1, \ldots, N\}$ be two MDCT bases of $\mathbb{R}^N$ with respectively long frame ℓ_{ton} to achieve good frequency resolution for tonals and short frame ℓ_{tran} to achieve good time resolution for transients. The MDCT is a bijective linear transform and we note $n_{ton} = \frac{N}{\ell_{ton}}$ and $n_{tran} = \frac{N}{\ell_{tran}}$ the number of frames for each basis (see Figure 2). Here, n and m are time-frequency indexes and will be denoted in the following $n = (q, \nu) \in [1, n_{ton}] \times [1, \ell_{ton}]$ or $m = (q, \nu) \in [1, n_{tran}] \times [1, \ell_{tran}]$.

The signal is decomposed as a linear combination of atoms of the two basis V and U that account for the tonal and transient layers plus a residual part that accounts for the noise and that is not sparse with respect to the two considered bases. We denote $D = V \cup U$ the dictionary made as the union of these two bases. D is overcomplete in $\mathbb{R}^N$, and any $x \in \mathbb{R}^N$ admits infinitely many expansions in the form:

$$x = \sum_{n \in I} \alpha_n v_n + \sum_{m \in I} \beta_m u_m + r \ , \qquad (1)$$

where $I = \{1, \ldots, N\}$, α_n and β_m are the expansion coefficients and r represents the noise term.

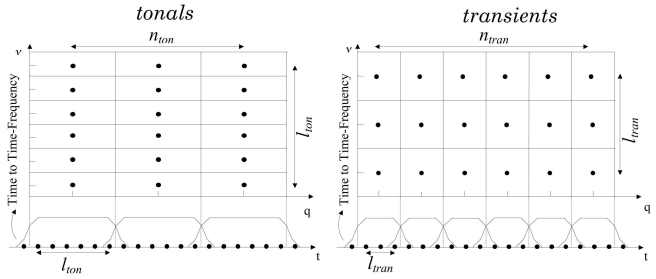

FIG. 2. Illustration of the two MDCT bases that account for the tonal part (long time resolution) and the transient part (short time resolution) of the signal.

We are interested in sparse signals, i.e. signals that may be written as:

$$x = \sum_{\lambda \in \Lambda} \alpha_\lambda v_\lambda + \sum_{\delta \in \Delta} \beta_\delta u_\delta + r \ , \qquad (2)$$

where Λ and Δ are small subsets of the index set $I = \{1, \ldots, N\}$ that account for the significant coefficients, i.e they identify which coefficient of the expansion are significantly non-zero. In what follows, they will be referred to as *significance maps*.

In order to model sparseness in the time-frequency coefficients, we introduce two indicator random variables corresponding to the significance maps Λ and Δ, $\gamma_{ton,n}$ and $\gamma_{tran,m} \in \{0, 1\}$ that control the sparsity of the expansion:

$$\gamma_{ton,n} = \begin{cases} 1 & \text{if } n \in \Lambda \\ 0 & \text{otherwise} \end{cases} \quad \gamma_{tran,m} = \begin{cases} 1 & \text{if } m \in \Delta \\ 0 & \text{otherwise} \ . \end{cases} \qquad (3)$$

We can therefore rewrite Eq. (2) as:

$$x = \sum_{n \in I} \gamma_{ton,n} \alpha_n v_n + \sum_{m \in I} \gamma_{tran,m} \beta_m u_m + r \ . \qquad (4)$$

The hybrid model is defined by two components: a discrete probability model for the significance maps, and a probability model for the expansion coefficients conditional upon the significance maps. Both of them are described below.

B. Coefficient priors

The sparseness of the expansion is conceptualized in a hierarchical manner as it is not directly modeled in the coefficients but through the binary indicator random variables $\gamma_{ton,n}$ and $\gamma_{tran,m}$ that are attached to each coefficient. As a result, hierarchical priors are given to the coefficients. We assume that, conditional upon the significance maps Λ and Δ , the coefficients α_n and β_m

are independent zero-mean normal random variables:

$$p(\alpha_n|\gamma_{ton,n},\sigma_{ton,n}) = (1-\gamma_{ton,n})\delta_0(\alpha_n) + \gamma_{ton,n}\mathcal{N}(\alpha_n|0,\sigma^2_{ton,n}) \quad (5)$$
$$p(\beta_m|\gamma_{tran,m},\sigma_{tran,m}) = (1-\gamma_{tran,m})\delta_0(\beta_m) + \gamma_{tran,m}\mathcal{N}(\beta_m|0,\sigma^2_{tran,m}) \ ,$$

where δ_0 is the Dirac delta distribution and, following (Wolfe *et al.*, 2004; Févotte *et al.*, 2008), the variances $\sigma_{ton,n}$ and $\sigma_{tran,m}$ are given an inverse-Gamma conjugate prior distribution:

$$p(\sigma^2_{ton,n}|\gamma_{ton,n}=1,f_{ton,n}) = \mathcal{IG}(\sigma^2_{ton,n}|1,f_{ton,n}) \quad (6)$$
$$p(\sigma^2_{tran,m}|\gamma_{tran,m}=1,f_{tran,m}) = \mathcal{IG}(\sigma^2_{tran,m}|1,f_{tran,m}) \ ,$$

where the scale parameters $f_{ton,n}$ and $f_{tran,m}$ are parametric frequency profiles that aim at taking into account the decrease of the energy of the signal when the frequency increases (Févotte *et al.*, 2008, Eq. (8)):

$$f_{ton,n} = \frac{\lambda_{ton}}{1+\left(\frac{q-1}{\ell_{ton}/3}\right)} \quad f_{tran,m} = \frac{\lambda_{tran}}{1+\left(\frac{q-1}{\ell_{tran}/3}\right)} \ . \quad (7)$$

The parameters λ_{ton} and λ_{tran} are given a non-informative *Gamma* conjugate prior.

Sparsity is enforced when $\gamma_{ton,n}=0$ (resp. $\gamma_{tran,m}=0$). In this case, the coefficients α_n (resp. β_m) are set to zero.

C. Indicator variable priors

In order to enforce structure between expansion coefficients, the significance maps Λ and Δ are given structured priors. We design priors that are tailored to the music signal. The one corresponding to the tonal basis encodes musical information based on harmonic content information using a chord transcription of the analyzed excerpt. The one corresponding to the transient basis is based on metrical content information using the sequence of beats corresponding to the analyzed excerpt.

To show the relevance of the proposed priors, the proposed approach will be systematically compared with a closely related state-of-the art approach described in Section II.C.3.

In what follows, some results will be illustrated through the representation of the significance maps that are defined by the two binary indicator random variables $\gamma_{ton,n}$ and $\gamma_{tran,m}$. In Figure 2, the MDCT bases are illustrated by a tiling of the time-frequency plane, where each tile represents a particular atom. The significance maps in the time-frequency plane are represented through the binary indicator random variable γ. To each atom corresponds an indicator variable that controls whether this atom is selected ($\gamma=1$) or not ($\gamma=0$)[2].

1. Model for tonals

For the significance map corresponding to the tonals, we propose to model dependencies between indicator variables using information about the harmonic content of the audio signal. Ideally, we would assume that we know the score corresponding to the musical excerpt and that, for each time frame $q \in \{1,\ldots,n_{ton}\}$, we know which notes the signal is composed of.

However, here, we want to work directly on audio, for which an exact transcription is usually not available. A number of recent work have shown that it is possible to accurately extract robust mid-level representation of the music, such as the chord progression (Papadopoulos and Peeters, 2011), that characterizes its harmonic content.

We propose to give a musical prior to the indicator variables using musical information obtained from a chord progression estimated from the audio file. The output of a chord estimation algorithm consists in a progression of chords chosen among a given chord lexicon that, in general, does not distinguish between any possible combination of simultaneous notes, but is typically reduced to a set of chords of 3 or 4 notes. The number of notes composing the chords will be denoted by N_c in the following. Here, we limit our chord lexicon to the 24 major and minor triads ($N_c=3$). The method we propose could be extended to larger dictionaries.

Each chord is characterized by a set of semitone pitch classes or chroma that correspond to the notes it is composed of. The chord progression does not provide an exact transcription of the music. For instance, passing notes are in general ignored, missing notes in the harmony may be added. Moreover, the chords are estimated regardless of octave. However, experiments show that the provided music content information is sufficient enough to build musically meaningful priors.

We consider two methods for building the structured significance maps for the tonal layer. In the first case, denoted as *Method Chord*, we use chord information. In the second case, denoted as *Method Chroma*, the priors are built relying directly on chromagram information.

a. Mapping between MDCT bins and chroma: In order to select atoms of the MDCT base that correspond to the harmonic content of the signal, we first perform a mapping between the MDCT bins and the 12 semitone pitch classes. Given a fixed frame index q, let $\{p^{MDCT}_\nu\}_{\nu=1,\ldots,\ell_{ton}}$ denote the semitone pitch classes corresponding to each frequency MDCT bin.

Assuming a perfect tuning of $A=440$Hz, a MDCT bin of frequency ν is converted to a chroma p^{MDCT}_ν by the following equation:

$$p^{MDCT}_\nu = (12\log_2\frac{\nu}{440}+69) \ (\text{mod } 12)[3] \ . \quad (8)$$

Note that, a single semitone pitch-class corresponds to several consecutive bins of the MDCT. Because of the logarithmic scale of Western tonal music, the higher the frequency, the larger the number of MDCT bins correspond to a single pitch class.

b. Method Chords: Given a fixed frame index q, let $\{p_k^{chord}\}_{k=1,\ldots,N_c}$ denote the semitone pitch-classes (chroma) corresponding to the estimated chord c_q. All bins of the MDCT that correspond to a note belonging to the estimated chord are selected. The indicator variables $\{\gamma_{ton,(q,\nu)}\}_{\nu=1,\ldots,\ell_{ton}}$ are given the following membership probabilities:

$$P_\Lambda\{\gamma_{ton,(q,\nu)}=1\} \quad (9)$$
$$= \begin{cases} p_{ton} & \text{if } \exists k\in[1,N_c] \mid p_\nu^{MDCT}=p_k^{chord} \\ 1-p_{ton} & \text{otherwise} , \end{cases}$$

where $0 \leq p_{ton} \leq 1$. The significance maps corresponding to the tonal layer should reflect the harmonic content of the audio signal. In practice, the value p_{ton} will be close to 1 (in our experiments, $p_{ton}=0.9$) so that atoms corresponding to the notes that are played are given high prior. The significant map for the tonal layer corresponding to the *Glockenspiel* monophonic audio signal of our test-set is illustrated in Figure 3. A set of atoms is selected at each frame according to the notes of the (note) transcription, regardless of octave. For instance all atoms $\{B1, B2, \ldots\}$ corresponding to the semitone B are selected when the first B note of the *Glockenspiel* is sounded.

The significance maps are given structures of "tubes" that have a musical meaning. Note also that we provide here a "vertical structure" for tonals.

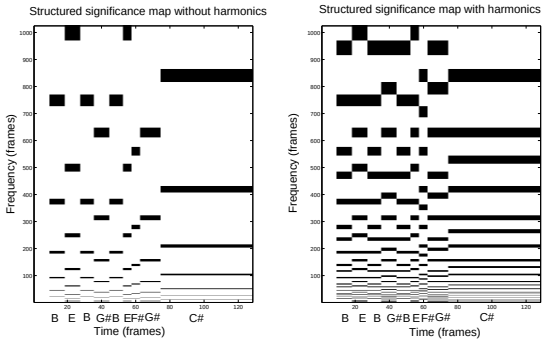

FIG. 3. Structured significance map for tonals for a *Glockenspiel* excerpt using chord information. Left: only notes composing the chords are considered. Right: higher harmonics are considered. The note transcription is indicated in the bottom.

c. Method Chroma: Given a fixed frame index q, let $\{a_k\}_{k=1,\ldots,12}$ denote the amplitude of each bin $\{p_k^{chroma}\}_{k=1,\ldots,12}$ of the computed chroma vector. Note that, according to (Papadopoulos and Peeters, 2011), the chromagram has been normalized so that $\sum_{k=1}^{12} a_k = 1$. For each chroma bin, all MDCT bins that correspond to this chroma are selected and given a weighted contribution according to the amplitude in the chroma vector. The indicator variables $\{\gamma_{ton,(q,\nu)}\}_{\nu=1,\ldots,\ell_{ton}}$ are given the following membership probabilities:

$$P_\Lambda\{\gamma_{ton,(q,\nu)}=1\} \quad (10)$$
$$= a_k \;\text{ if } \exists k\in[1,12] \mid p_\nu^{MDCT}=p_k^{chroma} .$$

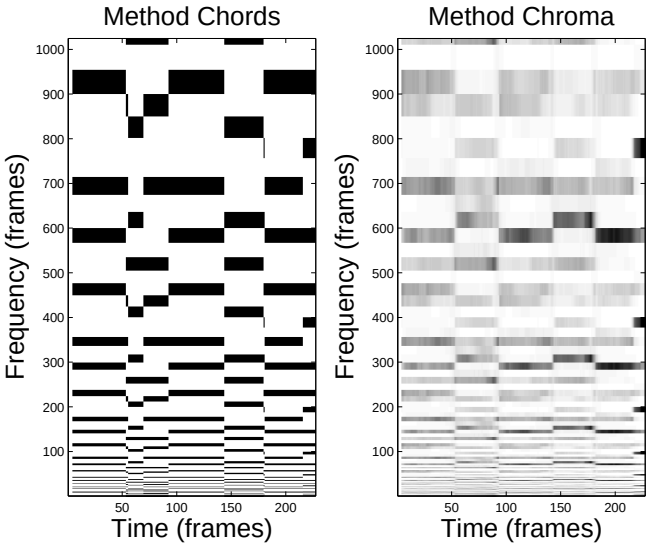

FIG. 4. Structured significance map for tonals for a *Beethoven String Quartet Op.127* excerpt using chroma information.

Figure 4 shows the significant maps for the tonal layer corresponding to the *Beethoven String Quartet Op.127* audio signal of our test-set obtained with *Method Chord* [left] and *Method Chroma* [right].

Two additional components may be added to improve the model.

d. Tuning: The instruments may have been tuned according to a reference pitch different from the standard $A4 = 440$Hz. In this case it is necessary to estimate the tuning of the track and Eq. (8) becomes:

$$p_\nu^{MDCT} = (12\log_2\frac{\nu}{A_{est}} + 69) \pmod{12} , \quad (11)$$

where A_{est} denotes the estimated tuning, here obtained with the method proposed in (Peeters, 2006).

e. Harmonics: Higher harmonics may be considered in the model. Each note produces a set of harmonics, whose frequencies are whole number multiples of the fundamental frequency[4], that results in a mixture of non-zero values in the chroma vector corresponding to the chord. For instance a C1 note will produce the set of harmonics $\{C1 - C2 - G2 - C3 - E3 - G3 - \ldots\}$. They can be considered in the significance maps, as illustrated in the right part of Figure 3. Here we take into account the first 6 harmonics of the notes[5]

2. Model for transients

For the significance map corresponding to the transients, we propose to model dependencies between indicator variables using information about the metrical structure of the audio signal. The idea is that, in a piece of music, most of the transient sounds will occur on beats or beat subdivisions. For instance, drum sounds are generally used to underline the metrical structure (beats, downbeats); in a string quartet piece, bow changes will generally occur on note changes, which are related to the metrical structure.

The structured prior corresponding to the significance map for tonals is built as follows. The beat positions are estimated from the signal using the beat tracker proposed in (Peeters and Papadopoulos, 2011), described in Section I.B.2. Let $\{b_k\}_{k=1,\ldots,N_b}$ denote the N_b beat positions (in frames) of the track (subdivisions of beats such as quarter notes or eighth notes can be considered as well). The indicator variables $\{\gamma_{tran,(q,\nu)}\}_{\nu=1,\ldots,\ell_{tran}}$ are given the following membership probabilities:

$$\begin{aligned} \forall \nu = 1,\ldots,\ell_{tran} \quad & P_\Delta\{\gamma_{tran,(q,\nu)} = 1\} \\ = & \begin{cases} p_{tran} & \text{if } \exists k \in [1, N_b] \mid q = k \\ 1 - p_{tran} & \text{otherwise} , \end{cases} \end{aligned} \tag{12}$$

where $0 \leq p_{tran} \leq 1$. The significance maps corresponding to the transient layer should reflect the metrical content of the audio signal. In practice, the value p_{tran} will be close to 1 (in our experiments, $p_{tran} = 0.9$) so that atoms corresponding to beat locations are given high prior. The significant map for the transient layer corresponding to the *Glockenspiel* audio signal of our test-set is illustrated in Figure 5. For each beat location, all frequency bins are retained, resulting in vertical lines that are sparse in time but cover all the frequencies, in the significance map. It may be noticed that the duration between two consecutive lines is not constant, as there my by variations in the tempo.

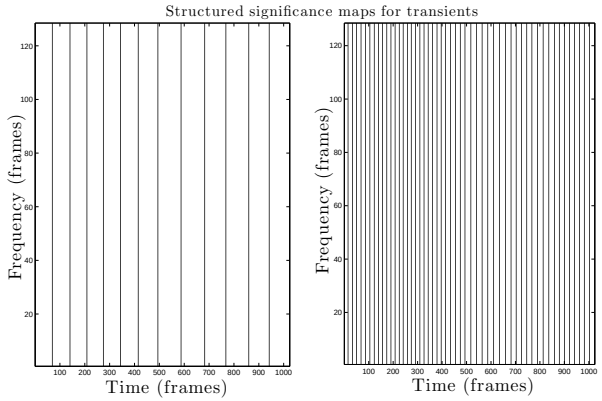

FIG. 5. Structured significance map for transients for a *Glockenspiel* excerpt using beat location information, resulting in vertical lines in the time-frequency plane. Left: considering beat locations. Right: considering eight-notes locations.

Remark 1 *As underlined in (Kowalski and Torrésani, 2008; Févotte* et al.*, 2008), the window lengths for each layer must be significantly different enough to discriminate between tonals and transients. Better results are obtained using a very short window length for the transients ($\approx 3ms$) that is shorter than the duration of a short attack. In this case, several frames may be needed to describe a transient. In practice, we also select vertical lines surrounding the theoretical (or estimated) beat locations in the transient layer prior. Eq.* (12) *thus becomes:*

$$\begin{aligned} \forall \nu = 1,\ldots,\ell_{tran} \quad & P_\Delta\{\gamma_{tran,(q,\nu)} = 1\} \\ = & \begin{cases} p_{tran} & \text{if } \exists k \in [1, N_b] \mid q = k \\ p_{tran} & \text{if } \exists k \in [1, N_b] \mid q = k-1 \\ p_{tran} & \text{if } \exists k \in [1, N_b] \mid q = k+1 \\ 1 - p_{tran} & \text{otherwise} . \end{cases} \end{aligned} \tag{13}$$

3. Baseline approach for comparison

In this article, we compare our approach, in which structural constraints on the coefficients are based on musical prior, to the baseline model (Févotte *et al.*, 2008), in which structural constraints on the coefficients relying on physical properties of the signal are imposed for both layers. Structure between significant coefficients of the decomposition is introduced to model persistence properties of time-frequency representations of audio signals, so that a horizontal prior structure is given to the indicator variables corresponding to the tonal layer, while vertical prior structure is given to the indicator variables corresponding to the transient layer.

For the tonal layer, persistency in time of the time-frequency coefficients is modeled. Given a fixed frequency index ν, the sequence $\{\gamma_{ton,(q,\nu)}\}_{q=1,\ldots,n_{ton}}$ is modeled by a two-state first-order Markov chain with transition probabilities $P_{ton,00}$ and $P_{ton,11}$, assumed equal for all frequency indices and initial distribution π_{ton} of each chain taken to be its stationary distribution.

For the transient layer, persistency in frequency of the time-frequency coefficients is modeled. Given a fixed frame index q, the sequence $\{\gamma_{tran,(q,\nu)}\}_{\nu=1,\ldots,\ell_{tran}}$ is modeled by a two-state first-order Markov chain with probabilities $P_{tran,00}$ and $P_{tran,11}$, assumed equal for all frames, and with learned initial probability π_{tran}.

The tonal and transient models are illustrated in Figure 6. We refer the reader to (Févotte *et al.*, 2008) for more details.

D. Residual

The residual signal r is modeled as an independent, identically distributed (i.i.d) Gaussian white noise, with variance σ^2, which is given an inverse-Gamma conjugate prior [6].

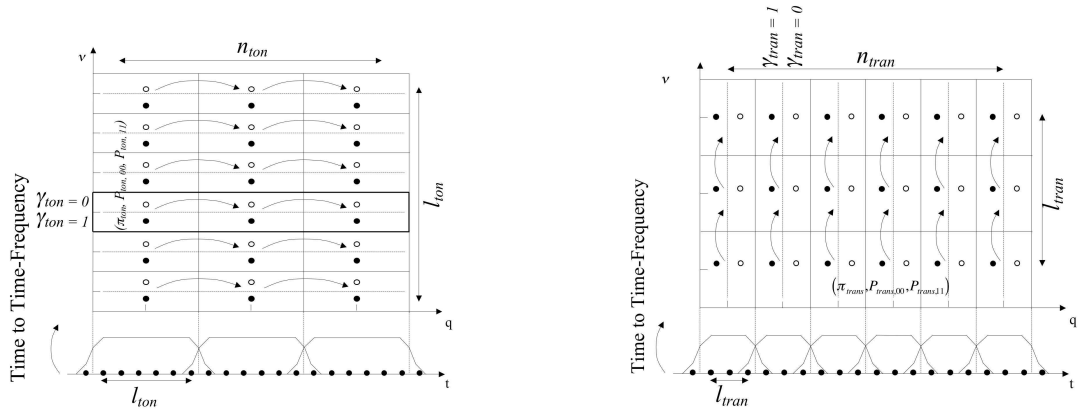

FIG. 6. Structured horizontal model for tonals and vertical model for transients based on time-frequency persistency properties. Adapted from (Févotte *et al.*, 2008).

III. MCMC INFERENCE

In the spirit of some previous work on Bayesian variable selection (Geweke, 1996; George and McCulloch, 1997), and following (Wolfe *et al.*, 2004; Févotte *et al.*, 2008), the posterior distribution of the set of parameters and hyperparameters of the model, denoted by $\theta = \{\alpha, \beta, \sigma_{ton}, \sigma_{tran}, \nu_{ton}, \nu_{tran}\} \cup \sigma$, is sampled from using a Gibbs sampler (Geman and Geman, 1984; Casella and George, 1992). Gibbs sampler is a standard Markov Chain Monte Carlo (MCMC) technique that simply requires to iteratively sample from the posterior distribution of each parameter, conditionally upon the data x and the remaining parameters.

The MCMC inference scheme we use is similar to the one described in (Wolfe *et al.*, 2004; Févotte and Godsill, 2006; Févotte *et al.*, 2008), the main difference being the new musical priors considered for the indicator variables we have introduced in Section II.C. For the sake of completeness, we provide in this section its general outline, using the notations we adopted in this paper. Further details about derivation of the expression for the update steps of the parameters, can be found in (Geweke, 1996; George and McCulloch, 1997; Wolfe *et al.*, 2004; Févotte *et al.*, 2008).

Let θ_{-k} denote the set of parameters in θ except the parameter k. Using Bayes'rule, the conditional distribution of each parameter k conditional upon the other parameters and the data can be written as:

$$p(k|\theta_{-k}, x) \propto p(x|\theta)p(\theta) \ . \tag{14}$$

The conditional distributions are thus proportional to the likelihood of the data times the priors on the parameters.

In order to avoid a nonconvergent Markov chain in the Gibbs sampler, (γ_{ton}, α) and (γ_{tran}, β) need to be sampled jointly (Geweke, 1996). As pointed out in (Févotte *et al.*, 2008), the structure of the dictionary $D = [VU]$ and the gaussian noise assumption allows alternative block sampling of (γ_{ton}, α) and (γ_{tran}, β), with the benefit of avoiding any matrix inversion at each iteration of the Gibbs sampler. Indeed, with the gaussian noise assumption, the likelihood of the observations can be written as:

$$p(x|\theta) = (2\pi\sigma^2)^{-N/2} \ exp(-\frac{1}{2\sigma^2} \parallel x - V\alpha - U\beta \parallel_2^2) \ . \tag{15}$$

Because the Euclidian norm is invariant under rotation, Eq. (15) can be written as:

$$\begin{aligned} p(x|\theta) &= (2\pi\sigma^2)^{-N/2} \ exp(-\tfrac{1}{2\sigma^2} \parallel \underbrace{U^T(x - V\alpha)}_{x_{tran|ton}} - \beta \parallel_2^2) \\ &= (2\pi\sigma^2)^{-N/2} \ exp(-\tfrac{1}{2\sigma^2} \parallel \underbrace{V^T(x - U\beta)}_{x_{ton|tran}} - \alpha \parallel_2^2) \ . \end{aligned} \tag{16}$$

According to Eq. (16), conditionally upon β (resp. α) and the other parameters, inferring α (resp. β) is thus a simple regression problem with data $x_{ton|tran}$ (resp. $x_{tran|ton}$), variable α (resp. β) modeled as i.i.d. conditionally upon γ_{ton} (resp. γ_{tran}), and i.i.d. noise, that does not require any matrix inversion.

Briefly, (γ_{ton}, α) and (γ_{tran}, β) are jointly sampled from by 1) sampling γ_{ton} (resp. γ_{tran}) from the posterior conditional distribution $p(\gamma_{ton,n}|\sigma_{ton,n}, \sigma, x_{ton|tran})$ (resp. $p(\gamma_{tran,m}|\sigma_{tran,m}, \sigma, x_{tran|ton})$), and 2) sampling α (resp. β) from the posterior conditional distribution $p(\alpha_n|\gamma_{ton,n}, \sigma_{ton,n}, \sigma, x_{ton|tran})$ (resp. $p(\beta_m|\gamma_{tran}, \sigma_{tran,m}, \sigma, x_{tran|ton})$). The detailed posterior distributions are given in Appendix A, and we refer the reader to (Févotte *et al.*, 2008; Geweke, 1996) for more details.

The posterior distribution of the other parameters σ_{ton}, σ_{tran}, ν_{ton} and ν_{tran} are easy to sample from since they have conjugate prior distributions and thus the corresponding posterior will have the same form.

The principal steps of the Gibbs sampler are summarized in Table I, where K is the total number of iterations

of the Gibbs sampler and K_{Burnin} is the burn-in length, corresponding to the number of iterations required before the Markov chain $\{\theta^{(1)}, \theta^{(2)}, \ldots\}$ reaches its stationary distribution. We provide the full posterior distributions in Appendix A.

TABLE I. Gibbs sampler steps for parameters inference.

Initialize $\theta^{(0)}$
for $k = 1 : K + K_{Burnin}$ **do**
Update tonals
Update γ_{ton} and α
$\gamma_{ton}^{(k)} \sim p(\gamma_{ton}|\sigma_{ton}^{(k-1)}, \sigma^{(k-1)}, x_{ton|tran}^{(k-1)})$ (Eq. (A1))
$\alpha^{(k)} \sim p(\alpha|\gamma_{ton}^{(k)}, \sigma_{ton}^{(k-1)}, \sigma^{(k-1)}, x_{ton|tran}^{(k-1)})$ (Eq. (A5))
Update hyperparameters
$\sigma_{ton}^{(k)} \sim p(\sigma_{ton}|\alpha^{(k)}, \lambda_{ton}^{(k-1)})$ (Eq. (A7))
$\lambda_{ton}^{(k)} \sim p(\lambda_{ton}|\sigma_{ton}^{(k)})$ (Eq. (A9))
Update transients
Update γ_{tran} and β
$\gamma_{tran}^{(k)} \sim p(\gamma_{tran}|\sigma_{tran}^{(k-1)}, \sigma^{(k-1)}, x_{tran|ton}^{(k-1)})$(Eq. (A3))
$\beta^{(k)} \sim p(\beta|\gamma_{tran}^{(k)}, \sigma_{tran}^{(k-1)}, \sigma^{(k-1)}, x_{tran|ton}^{(k-1)})$ (Eq. (A6))
Update hyperparameters
$\sigma_{tran}^{(k)} \sim p(\sigma_{tran}|\beta^{(k)}, \lambda_{tran}^{(k-1)})$ (Eq. (A8))
$\lambda_{tran}^{(k)} \sim p(\lambda_{tran}|\sigma_{tran}^{(k)})$ (Eq. (A10))
Update noise
$\sigma^{(k)} \sim p(\sigma|\alpha^{(k)}, \beta^{(k)}, x)$ (Eq. (A11))
end for

The Minimum Mean Square Estimates (MMSE) of the parameters θ can then be computed from the Gibbs samples $\{\theta^{(K_{Burnin})}, \theta^{(K_{Burnin}+1)}, \ldots, \theta^{(K)}\}$ of the posterior distribution $p(\theta|x)$:

$$\hat{\theta}_{MMSE} = \int \theta p(\theta|x) d\theta \quad (17)$$

$$\approx \frac{1}{K - K_{Burnin}} \sum_{k=K_{Burnin}+1}^{K} \theta^{(k)} \; . \quad (18)$$

Time-domain source estimates are reconstructed by inverse transform of the estimated coefficients (inverse MDCT in our case). The denoised estimation is constructed by $\hat{x} = \alpha V + \beta U$.

IV. EXPERIMENTS PROTOCOL AND EVALUATION MEASURES

In this Section, we describe the test-set and measures we use for the evaluation of the proposed model.

A. Experimental setup

We present simulation results on 5 musical excerpts of various music styles that are described in Table II. These signals have been chosen because they have diverse characteristics: the *Glockenspiel* signal is a monophonic signal of a tuned percussion instrument, the *Misery* and *Love Me Do* signals are two complex polyphonic excerpts of Beatles songs containing voice and drum sounds, the *Beethoven String Quartet* signal is an excerpt of a string quartet and the *Mozart Piano Sonata* signal is an excerpt of polyphonic piano music.

TABLE II. Sound excerpts used for evaluation of the model. SR: sampling rate.

Name	SR (Hz)	Duration
Glockenspiel	44100	2s
Misery (Beatles)	11025	11s
Love Me Do (Beatles)	11025	5s
Beethoven String Quartet Op.127 – 1	11025	11s
Mozart Piano Sonata KV310 – 1	11025	11s

a. Parameters : The length of the two MDCT bases are set to 1024 samples for the tonal layer and 128 samples for the transient layer, at a sampling rate of 44100Hz, and respectively to 256 and 32 samples at a sampling rate of 11025Hz. The MDCT of the clean and noisy signals, with input $SNR = 10dB$ are represented in Figure 7. The MMSE and MAP estimates of the parameters are computed by averaging the last 100 samples of the Gibbs sampler, run for 500 iterations.

We compare a semi-automatic model (denoted as version *SA*), assuming that the transcription is known (notes for the monophonic signal, chords for the polyphonic signals, and beat locations) with a fully-automatic model where the music content is directly estimated from the input signal (denoted as version *A*). Our approach that incorporates musical priors is compared with the one presented in (Févotte *et al.*, 2008), described in Section II.C.3, in which the priors for both the tonal and transient layers are based on time-frequency persistency properties (version (Févotte *et al.*, 2008)).

B. Evaluation measures

1. Audio denoising:

In the context of audio denoising task, artificial noisy signals are created by adding Gaussian white noise to the clean signal with various input SNRs. The case without additional noise WN (without noise) corresponds to a separation into two layers *transient + tonal*. Partials are expected to be recovered in the tonal layer while attacks or percussive sounds will be recovered in the transient layer. The results in terms of output SNR provide an objective evaluation measure. However, although widely used for assessing algorithm performances, the SNR is not a completely relevant measure of distortion for audio signals and is insufficient to measure the quality or intelligibility of the reconstructed signals. Indeed, it does not reflect exactly the perceived audio quality that includes distortion of the reconstructed signal, musical noise or other artifacts. Subjective evaluation by listening to the signals is also required. Subjective quality assessment of signal reconstruction in source separation or denoising tasks is an active ongoing topic of re-

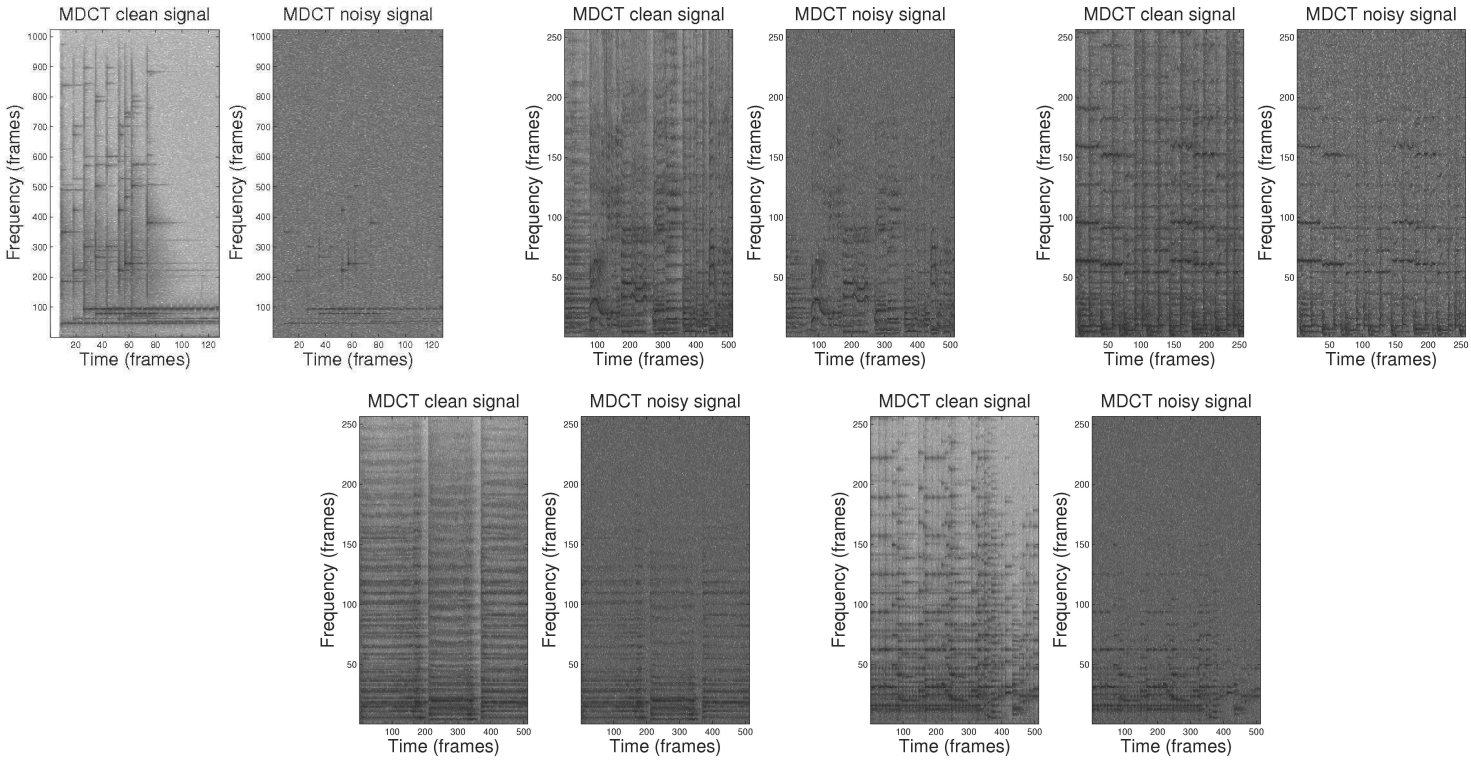

FIG. 7. MDCT of the clean and noisy signals, with input $SNR = 10dB$. Form top top bottom : Glockenspiel, Misery, Love Me Do, Beethoven and Mozart excerpts.

search (Vincent *et al.*, 2006; Emiya *et al.*, 2010; Rohdenburg *et al.*, 2005). It has been found that statistically significant results can be obtained from listening tests with less than ten non-expert subjects (Vincent *et al.*, 2006). However, conducting large-scale listening tests is out of scope of this work, and the subjective assessment of the results is limited here to an analysis and description by the authors of the audio excerpts obtained by the proposed algorithm. All the audio excerpts are available at `http://webpages.lss.supelec.fr/perso/matthieu.kowalski/jasa/jasa.htm` (date last viewed 01/20/13).

2. Sparsity:

In this work, we aim at obtaining a dual representation of the signal that is sparse, but that is also structured and is meaningful according to the music content of the analyzed audio excerpt. A number of criteria for measuring the sparsity of an expansion have been proposed. Among them, Rényi entropies (Rényi, 1961), a generalization of the Shannon entropy, have been introduced in (Baraniuk *et al.*, 2001) as measures for estimating the complexity and information content of a signal through its time-frequency representation. It has been shown that minimizing the complexity or information of a time-frequency representation of a signal is equivalent to maximizing its concentration, peakiness, and resolution. Let $\Phi \in \mathcal{L}^2(\mathbb{R}^2)$ be a time-frequency representation of a unit-energy signal $s \in \mathcal{L}^2(\mathbb{R})$ (for instance, in this article Φ is the MDCT transform of the signal).The Rényi entropies of Φ, defined as:

$$R_\alpha(\Phi(t,f)_s) = \frac{1}{1-\alpha}\ log_2 \int\int\ \Phi(t,f)_s^\alpha\ dtdf,\ \alpha \in [0,1]\ , \tag{19}$$

may be interpreted as sparsity measures and have thus been used as a criterion for obtaining a sparse expansion of an audio signal (Jaillet and Torrésani, 2004; Liuni *et al.*, 2011). In the present work, we use a Rényi entropy criteria as an evaluation measure to compare the sparsity of the significance maps. We present here results with $\alpha = 0.9$, similar results were obtained with different criteria.

Sparsity is also measured in terms of the percentage of atoms selected in the significance maps.

C. Computational performances

The algorithms are implemented in MATLAB and performed on a MacBook Pro Intel Core 2 Duo clocked at 2.4GHz with 2GB RAM. The computation time of the proposed method is similar to the one obtained with (Févotte *et al.*, 2008), ≈ 270 s for processing the *Glockenspiel* signal for instance. Note that the use of MCMC schemes generates high computational costs.

V. RESULTS AND DISCUSSION

The aim of this section is to provide an analysis of the proposed approach for sparse and structured expansion of audio signals on overcomplete dictionaries. In order to compare our approach to previous work, we evaluate the performance of the proposed approach for the task of audio denoising. Our purpose is to show how expert music knowledge can be used to build priors to obtain a relevant signal decomposition. The main contribution of the article is the design of new musical priors. In order to evaluate the impact of the prior itself, we compare our results with (Févotte *et al.*, 2008), that specifically differs from the proposed method in the priors. The impact of the various parameters (tuning, harmonics, and priors settings) is studied. We also provide a detailed analysis of our model in terms of sparsity and in terms of interpretability of the representation. We wish to show that the use of priors built on music content information allows making the structure of the signal legible and leads to a representation that has a musical/physical meaning.

A. Structured representation

Our aim here is to provide a structured representation of the signal that is meaningful from a musical point of view, in the sense that it highlights characteristics of the signal that are of interest. Such a relevant representation may be very useful for extracting higher-level information for a given MIR application. For instance, having a clear representation of the partials of the notes may be useful for removing the drum from a polyphonic excerpt.

The use of musical priors yields a structure that better reflects the music content of the signal compared to the approach that uses physical priors. Figure 8 shows the significance maps of the selected atoms (MMSE estimates) for the *Glockenspiel* signal, in the WN case, obtained with the (Févotte *et al.*, 2008) approach [left-top], using musical priors for the transient layer [left-bottom], using musical priors for the tonal layer [right-top], using musical priors for both the tonal and transient layers [right-bottom]. When using musical priors for the tonal layer, the resolution of the tonal significance map is sharper. The partials of the notes clearly appear as thin horizontal lines and are better discriminated, especially in low frequencies. The beginning of the notes is also clearer. The structure of the significance maps is even sharper and legible when we also use musical priors for the transient layer.

However, it should be noticed that, especially under low-input SNRs conditions, one may perceive artifacts in the reconstructed signal with the method we propose. These artifacts may be due, on the one hand, to the fact that some high frequencies are captured by the transient basis rather than by the tonal basis. On the other hand, the construction of the significance map corresponding to the transient layer leads to some regular "clicks" that are audible in low frequencies. But, in spite of these artifacts, one can find by listening to the reconstructed signals that results obtained relying on musical priors are generally "richer" than the ones obtained with the approach used in (Févotte *et al.*, 2008). This is very clear for instance in the case of the *Beethoven* excerpt, with $SNR_{in} = 10$dB. The chords (especially those on beats 2 and 3) sound more "round", and the notes are better held. It should be noticed that listening tests in general do not reveal noticeable improvement when using musical prior also for the transient layer, even though the reconstructed signal in the case of the *Mozart* excerpt with $SNR_{in} =$ 0db sounds a little "flatter" without musical prior for the transient layer.

B. Semi-automatic versus automatic approach

1. Tonal layer

Table III shows that the SNR results obtained with an automatic approach (A), where the tonal content is directly estimated from the audio (using a chord estimation algorithm), are similar to the ones obtained with a semi-automatic approach (SA), where the transcription (chord ground-truth) is given as an input of the model. Results are similar at the same time in terms of SNR, in terms of legibility of the significance maps, and in terms of listening tests. The results of chord estimation are indicated in Table III and show that the chord estimation is not perfect. Two scores are considered: *Exact Estimation EE* corresponds to the rate of chords correctly detected; *Exact + Neighbor E+N* corresponds to the rate of correctly detected chords including neighboring chords [7]. It can be seen that the exact chord estimation results may be low, especially in low input SNR conditions (for instance $EE = 28.35\%$ for the *Love Me Do* excerpt, with $SNR_{in} = 0$. However, most errors correspond to neighboring chords (high $E + N$ results), which indicates that most of the notes present in the signal are correctly taken into account. A rough estimation of the tonal content is thus sufficient to built relevant prior for the significance map corresponding to the tonal layer.

2. Transient layer

The difference between the SNR results obtained with an automatic approach, where the beats are directly estimated from the audio (using a beat tracking algorithm) and a semi-automatic approach where the beat locations are given as an input of the model, was in each case ≤ 0.1dB (For conciseness, detailed results are not reported here). Here again, the transcription does not need to be perfect to build relevant prior for the significance map corresponding to the transient layer.

C. Influence of musical priors for the transient layer only

Table IV shows the effect of using musical priors for the transient layer only (and using frequency persistency-based priors for the tonal layer, as in (Févotte *et al.*, 2008)), denoted as case *Method tr.* In terms of SNR,

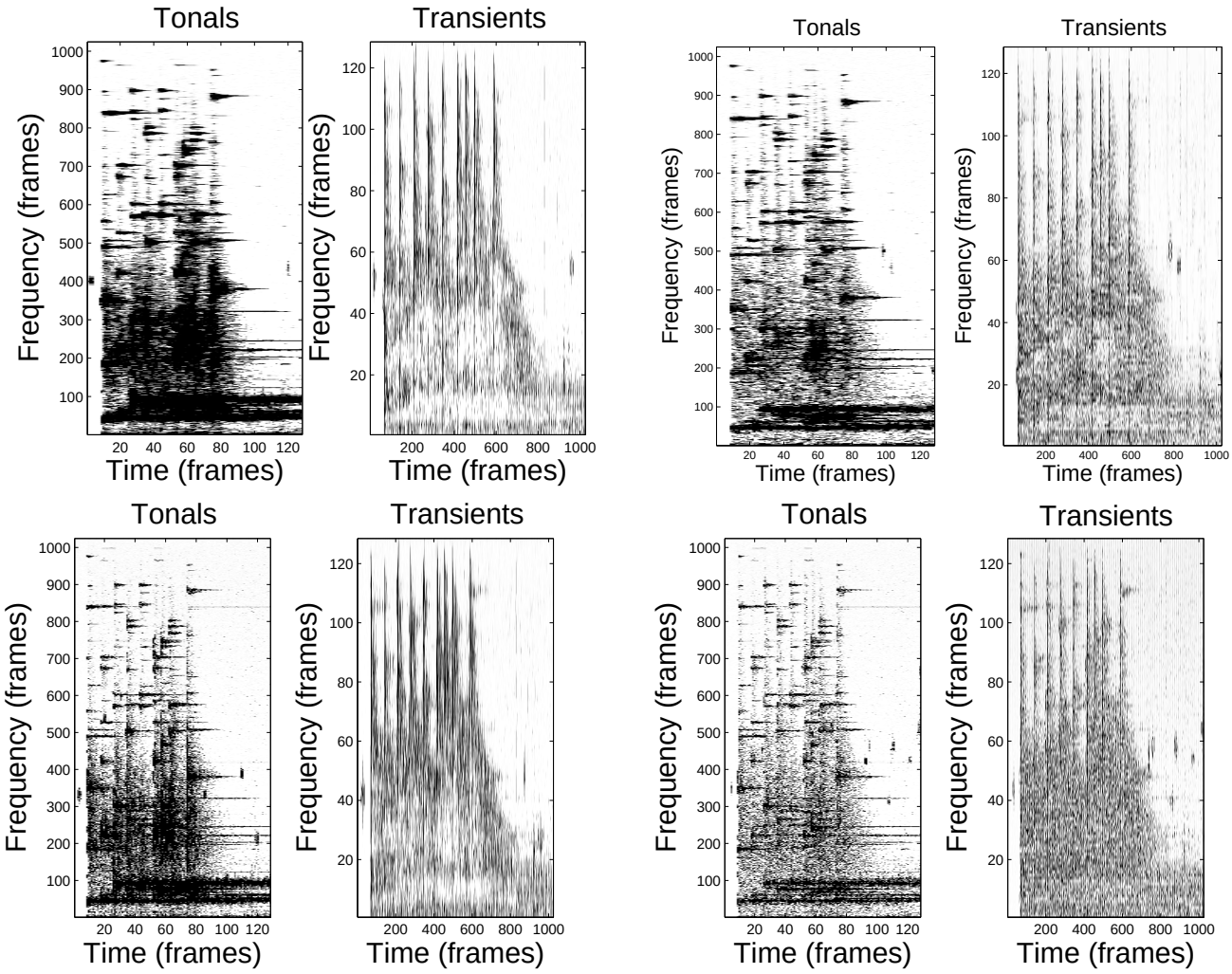

FIG. 8. Significance maps of the tonal and transient bases (MMSE estimates) for the *Glockenspiel* excerpt, case WN. Left-top: approach (Févotte *et al.*, 2008); Left-bottom: proposed approach using musical priors for the transient layer only; Right-top: proposed approach using musical priors for the tonal layer; Right-bottom: proposed approach using musical priors for both the tonal and transient layers.

musical priors for the transient layer yield results that are equivalent to state-of-the-art results, except for polyphonic clean signals. For monophonic and polyphonic noisy signals, the use of musical priors for transients results in a tonal layer that may be more consistent with the music content: the partials of the notes are more distinguishable, as illustrated in Figure 8 [Left-bottom]. However, for clean polyphonic signals, the use of beat positions as prior information for building the transit layer is too restrictive for having a satisfying decomposition. When listening to the results on the *Mozart* excerpt, one can hear that the residual part obtained with *Method tr.* is non-negligible compared to the one obtained with *Method (Févotte* et al.*, 2008)*, although it should be zero. For the Beethoven excerpt, some high frequencies are captured by the transient basis rather than by the tonal basis, as it is illustrated Figure 9. Using musical prior for transients based on the metrical structure thus shows some potential, as it is adapted to the semantic content of the signal, but it should be refined, for instance by using onset positions instead of beat positions. This is also discussed in Section V.E.

D. Comparison between chromagram versus chords

We have proposed two methods for building priors for the significance map corresponding to the tonal layer.

1. In the first case (*Method Chord*), the map is built using information about the tonal content from the estimated chords.

2. In the second case (*Method Chroma*), the map is built using information about the tonal content directly from the chromagram.

Table V shows that *Method Chord* outperforms *Method Chroma* in terms of SNR in the case without noise added

TABLE III. SNR results (in dB), and chord estimation results (*EE*: *Exact Estimation*, *E+N*: *Exact + Neighbor*) for various input SNRs and without additional Gaussian noise (WN), for comparison between a semi-automatic (estimated chords) and an automatic (given chords) approach.

	SNR_{in}	WN		0		10		20	
	Approach	*A*	*SA*	*A*	*SA*	*A*	*SA*	*A*	*SA*
Gl.	*EE*	14.39		14.39		14.39		14.39	
	$E+N$	72.27		72.27		72.27		72.27	
	SNR_{out}	71.07	71.36	14.15	14.13	21.31	21.37	28.58	28.59
Mi.	*EE*	79.22		79.02		79.22		78.77	
	$E+N$	82.00		88.1		82.00		81.54	
	SNR_{out}	42.28	42.73	6.36	6.35	13.02	13.02	20.91	20.87
Lo.	*EE*	95.29		28.35		60.58		96.27	
	$E+N$	100		92.62		100		100	
	SNR_{out}	29.12	28.31	6.44	6.42	12.44	12.44	19.24	19.21
Be.	*EE*	87.22		86.56		86.77		86.77	
	$E+N$	100		100		100		100	
	SNR_{out}	55.15	54.72	7.17	7.15	13.56	13.54	21.63	21.62
Mo.	*EE*	75.69		70.26		75.25		75.69	
	$E+N$	90.26		88.70		90.26		90.26	
	SNR_{out}	62.13	62.33	8.20	8.19	15.47	15.47	23.40	23.42

TABLE IV. SNRs results (in dB), for various input SNRs and without additional Gaussian noise (WN), for comparison between using musical priors for the transient layer only, case *tr*, and with the baseline approach (Févotte *et al.*, 2008) (*F2008*).

SNR_{in}	WN		0		10		20	
Method	*tr*	*F2008*	*tr*	*F2008*	*tr*	*F2008*	*tr*	*F2008*
Gl.	70.01	70.22	15.54	15.74	22.14	22.45	29.16	29.22
Mi.	33.38	44.41	6.89	6.90	13.13	13.29	20.43	21.08
Lo.	27.30	29.61	6.72	6.77	12.77	12.72	19.49	19.35
Be.	39.60	54.64	7.71	7.71	13.90	14.03	21.44	21.99
Mo.	54.47	60.96	8.98	8.97	15.87	15.94	23.59	23.88

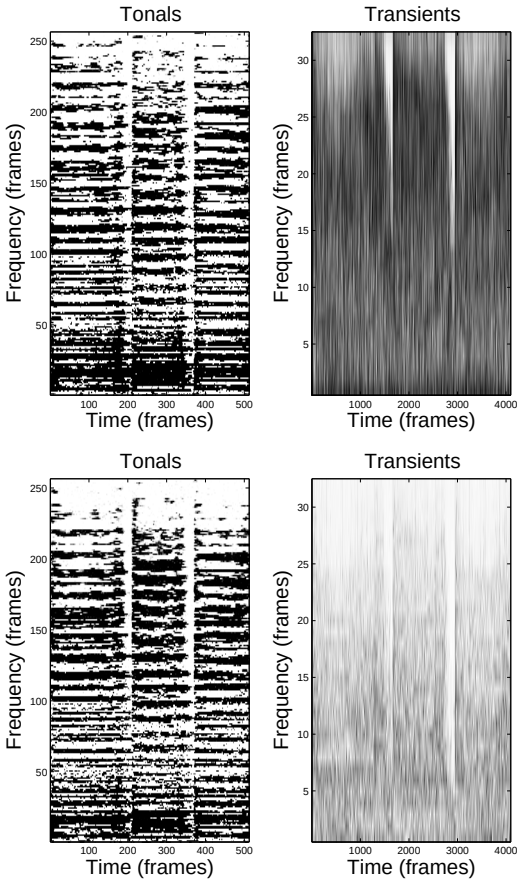

FIG. 9. Significance maps of the tonal and transient bases (MMSE estimates) for the *Beethoven* excerpt, case WN. Top: approach (Févotte *et al.*, 2008); Bottom: proposed approach using musical priors for the transient layer only.

to the clean signal, and in the case of the monophonic signal. *Method Chroma* slightly outperforms *Method Chord* in terms of SNR in the case of polyphonic music and when noise is added to the clean signal. Experiments reveal that the significance maps obtained with *Method Chroma* are sharper and sparser than those obtained with *Method Chord*, as illustrated in Figure 10. Moreover, listening tests reveal that the noise induced by the proposed method in the case of *Method Chord*, is considerably reduced when using *Method Chroma*, especially when musical priors are used both for the tonal and the transient layers (see for instance the *Beethoven* excerpt, case $SNR_{in} = 10$dB).

However, reconstructed signals using *Method Chroma* are not systematically more pleasant to listen too. A drawback of *Method Chroma* is that, because the hierarchy between the 12 pitch classes is less strong than with *Method Chord* when building the prior for the tonal layer (because all pitch classes are given a non-zero contribution, whereas in the case of *Method Chord*, only 3 pitch classes are considered at each time-instant), the reconstructed signal may select notes that do not fit the harmonic content, similarly to the one obtained with approach (Févotte *et al.*, 2008). In this case, one may perceive "wrong" notes in the reconstructed signal, especially in high frequencies, as it can be heard, for instance, when listening to the *Mozart* excerpt, case $SNR_{in} = 10$dB. This blurry phenomenon has especially an impact in the case on clean and monophonic signals.

TABLE V. SNRs results (in dB), for various input SNRs and without additional Gaussian noise (WN), for comparison between *Method Chord* (*Cho*) and *Method Chroma* (*Chr*).

SNR_{in}	WN		0		10		20	
Method	*Cho*	*Chr*	*Cho*	*Chr*	*Cho*	*Chr*	*Cho*	*Chr*
Gl.	71.35	69.90	14.13	13.27	21.37	20.70	28.59	27.87
Mi.	42.73	37.09	6.35	7.04	13.02	13.35	20.87	20.89
Lo.	28.32	26.19	6.42	6.80	12.44	12.85	19.21	19.27
Be.	54.72	54.62	7.15	8.63	13.53	14.52	21.62	22.05
Mo.	62.33	60.1	8.19	9.45	15.47	16.28	23.42	23.96

TABLE VI. SNR results for comparison between the proposed method that uses musical priors and the approach proposed in (Févotte *et al.*, 2008) (*F2008*). Case *A*: musical priors only for the tonal layer (automatic approach). Case *A + tr*: musical priors for the tonal and the transient layer.

SNR_{in}	WN			0			10			20		
Method	A	$A+tr$	*F2008*	A	$A+tr$	*F2008*	A	$A+tr$	*F2008*	A	$A+tr$	*F2008*
Gl.	71.35	71.49	70.22	14.13	13.86	15.74	21.37	20.98	22.45	28.59	28.24	29.22
Mi.	42.73	32.72	44.41	7.03	7.04	6.9	13.35	13.34	13.29	20.89	20.60	21.08
Lo.	28.32	27.14	29.61	6.80	6.72	6.77	12.85	12.95	12.72	19.27	19.72	19.35
Be.	54.72	35.97	54.64	8.63	8.65	7.71	14.52	14.54	14.03	22.05	21.49	21.99
Mo.	62.33	57.07	60.96	9.45	9.37	8.97	16.28	16.12	15.94	23.96	23.73	23.88

E. Denoising quality

Table VI compares results obtained with the proposed method that uses musical priors with the approach proposed in (Févotte *et al.*, 2008). As underlined in the previous section, *Method Chroma* has a negative "blurry" effect on clean and monophonic signals. Ideally, the method should be selected according to the type of test signal. For the sake of legibility, we do not report results with both methods in Table Table VI. Our purpose here is to demonstrate the potential of the proposed musical priors, compared to existing approaches. We thus present the best results, obtained with *Method Chroma* in the case of polyphonic music and when noise is added to the clean signal and *Method Chord* otherwise. Concerning the quality of denoising, our model provides results that are comparable to state-of-the-art algorithms in terms of SNR: the difference between our method and method (Févotte *et al.*, 2008) are in general lower than 1 dB. The method proposed in (Févotte *et al.*, 2008) slightly outperforms our method in the case of monophonic music, whereas our method performs slightly better in the case of polyphonic music. Our explanation is that, by relying on chord information, too many notes are considered when building the prior corresponding to the tonal significance map of the monophonic excerpt. A polyphonic/monophonic detection step could be added to improve the proposed model.

Differences between the two approaches may be perceived while listening to the sound files. As said before, the proposed approach induces some artifacts. In particular, the quality of denoising obtained with the proposed prior for the transient layer is disappointing. Indeed, the construction of the significance map corresponding to the transient layer leads to some regular "click" sounds that are superimposed to the signal of interest. As indicated in Table VII and illustrated in Figure 10 and 11, compared to (Févotte *et al.*, 2008), our approach selects much less atoms in the transient layer. Although the idea seems "natural", the use of beat position information for building the prior for the transient layer may be too restrictive for denoising purpose.

When we build the tonal significance map, we select at each time instant all the MDCT bins that correspond to the notes of the estimated chord, regardless of octaves. Because the estimation of the tonal content is rough (we do not have an exact transcription and ignore for instance passing tones), some of the "tubes" that result from this selection in the tonal significance map may correspond to added notes that are not actually in the signal and thus not completely relevant. However, these atoms are musically coherent with the music content (they are coherent with the underlying harmony) and are not as disturbing as if they were atoms randomly selected. Figure 11 shows the significance maps (MMSE estimates) for the *Mozart* signal, in the case $SNR_{in} = 10$dB. It can be seen that in the case of (Févotte *et al.*, 2008) approach, many atoms that do not fit the tonal content are selected, especially in high frequencies. When listening to the reconstructed signal, one may perceive some notes that seem "wrong". When listening to the signal obtained with the proposed approach, one may also perceive somme added notes, but they are coherent with the harmonic content. This is important when extracting higher-level information from the reconstructed signal. For instance, in the case of key estimation, the algorithm will be less affected by notes that belong to the tonal content than by notes that seem randomly selected.

F. Sparsity

In the case of polyphonic music, the proposed approach provides a representation that is sparser than the one proposed in (Févotte *et al.*, 2008). Indeed, the number of remaining non-zero coefficients in contrast to the total numbers of atoms of the initial redundant dictionary is usually lower when using musical priors, as well as Renyi entropy, as it can be seen in Table VII. Figures 8, 10 and 11 illustrate the fact that the energy density of the significance map corresponding to the tonal layer is concentrated into thin horizontal lines, and Figures 8, 10 and 12 clearly illustrate that the energy density of the significance map corresponding to the transient layer is is concentrated into thin vertical lines. In the case of monophonic music, figures in Table VII indicate that our representation is less sparse than the one provided in (Févotte *et al.*, 2008). This is because our method induces some artifacts resulting from undesirable selected atoms (see Figure 12, [middle]). The representation may become sparser by increasing the value of p_{ton} in Eq. (9), as explained in Section V.G.1. Note that, as illustrated in Figure 8, in the case of running the algorithm on a clean signal (case WN), our approach provides a representation that is sparser than the one obtained with (Févotte *et al.*, 2008), where a small noise is needed to obtain a

TABLE VII. Percentage of remaining non-zero coefficients selected for the tonal layer (% *ton.*) and the transient layer (% *tran.*) and value of Renyi entropy for tonal layer (*Renyi ton.*) and the transient layer (*Renyi tran.*) (MAP estimates). Case: *F2008*: approach in (Févotte *et al.*, 2008). Case *A*: musical priors only for the tonal layer (using estimated chords). Case *A + tr*: musical priors for the tonal and the transient layer.

	SNR_{in}	WN			0			10			20		
	Method	A	$A+tr$	*F2008*	A	$A+tr$	*F2008*	A	$A+tr$	*F2008*	A	$A+tr$	*F2008*
Gl.	% *ton.*	26.1482	19.7510	30.3772	2.0317	1.9966	1.5022	2.5848	2.4879	1.5518	3.3974	3.2333	3.3745
	% *tran.*	29.0039	33.0078	24.6094	0	8.9844	0	0	6.8359	0	0	5.9570	0
	Renyi ton.	15.0604	14.6550	15.2776	11.3662	11.3403	10.9392	11.7111	11.6559	10.9812	12.1047	12.0313	12.0875
	Renyi tran.	15.2006	15.3874	14.9631	NaN	13.5122	NaN	NaN	13.1180	NaN	NaN	12.9194	NaN
Mi.	% *ton.*	23.1979	14.7957	50.7500	3.0151	3.0815	15.7799	5.2574	5.5031	25.5539	9.0256	9.4261	37.4146
	% *tran.*	58.3740	32.6172	44.8242	11.8164	16.2109	12.3047	21.5820	19.3359	48.9014	46.1914	23.3398	67.8223
	Renyi ton.	14.8886	14.2306	16.0161	11.9396	11.9697	14.3331	12.7418	12.8048	15.0260	13.5188	13.5808	15.5774
	Renyi tran.	16.1770	15.3702	15.7958	13.9016	14.3637	13.9315	14.7733	14.6180	15.9218	15.8713	14.8896	16.3935
Lo.	% *ton.*	16.6702	12.2955	45.8679	2.6535	2.8152	11.4395	4.8920	5.0903	20.6650	8.7173	8.7738	32.9956
	% *tran.*	74.0723	37.5000	27.0020	31.2500	20.3125	10.2051	45.7031	26.5625	19.4824	51.9531	32.2266	24.6582
	Renyi ton.	13.4112	12.9614	14.8705	10.7474	10.8349	12.8587	11.6348	11.6909	13.7200	12.4686	12.4759	14.3943
	Renyi tran.	15.5207	14.5736	14.0640	14.3106	13.6891	12.6616	14.8560	14.0761	13.5926	15.0413	14.3550	13.9329
Be.	% *ton.*	30.2101	19.2009	44.4839	3.1097	3.0518	8.9005	5.6511	5.9242	18.5188	9.5741	10.3973	32.7576
	% *tran.*	57.5195	31.2500	61.3037	2.9297	6.8359	13.4033	23.4375	6.7383	82.7148	80.0781	15.8203	99.8779
	Renyi ton.	15.2689	14.6063	15.8275	11.9828	11.9540	13.5067	12.8449	12.9090	14.5619	13.6028	13.7216	15.3836
	Renyi tran.	16.1560	15.3084	16.2482	11.8713	13.1180	14.0549	14.8926	13.0972	16.6802	16.6664	14.3285	16.9521
Mo.	% *ton.*	36.0977	21.1670	42.2409	2.9648	2.9938	8.1978	4.6402	4.8706	12.5305	7.4661	7.5645	18.7614
	% *tran.*	39.4287	54.6875	33.2764	4.1992	14.7461	14.4043	39.3555	15.8203	47.2656	79.9805	28.8086	92.5049
	Renyi ton.	15.5174	14.7432	15.7447	11.9101	11.9226	13.3835	12.5500	12.6164	13.9843	13.2364	13.2536	14.5687
	Renyi tran.	15.6106	16.1180	15.3662	12.4149	14.2271	14.1588	15.6433	14.3285	15.8731	16.6655	15.1933	16.8414

good decomposition.

G. Impact of parameters

1. Indicator variable prior set-up:

The values p_{ton} in Eq. (9) and Eq. (10) and p_{tran} in Eq. (13) have an effect on the above-mentioned artifacts produced by our model in low-input SNRs conditions. For instance, setting p_{ton} and p_{tran} to 0.99 instead of 0.9 in the case of the *Glockenspiel* signal allows reducing the artifacts for $SNR_{in} = 10$dB, as it can be seen in Figure 12. However, our experiments show that indicator variables corresponding to atoms that do not belong to the chord must not be set to 0. Setting p_{ton} or p_{tran} to 1 results in reconstructed signals of very "poor" sound, as it can be assessed by listening tests. Output SNRs are also degraded. Setting $p_{ton} < 1$ allows taking into account imperfections of the chromagram given as input of the hybrid model (temporal imperfections due to windowing, discrepancy between the ideal model and reality, *etc.*).

2. Impact of tuning:

Integrating tuning information in the model does not lead to improvement in terms of output SNR values (detailed results are not reported here to avoid overfill of the article), but yields to estimated significance maps that are more coherent with our model. Indeed, the "tubes" depend on the tuning and thus, in case of "bad" tuning, using tuning information allows selecting atoms within the correct frequency regions.

3. Impact of harmonics:

We did not find any improvement when adding harmonics in our model: the difference between the SNR results obtained with and without considering the harmonics was systematically $\leq 0.3dB$ (detailed results are not reported here to avoid overfill of the article). This may be explained by the fact that, in the polyphonic case, the contribution of a large part of the first 6 higher harmonics of a note is already taken into account in the significance map by the other notes. For instance, let us consider C major chord (C-E-G). The C note generates harmonics E and G. E and G are thus both actual played notes and harmonics. Their contribution is already partially taken into account in the significance map in the case of the model "without harmonics" .

VI. CONCLUSION AND FUTURE WORK

In this article, we have presented a new approach for sparse decomposition of audio signals of music on an overcomplete dictionary made as the union of two MDCT bases. The originality of our model is that, within a Bayesian framework, we introduce musical priors that aim at modeling dependencies between the coefficients of the expansion in a more realistic way than what has been proposed before. The main contribution of the article is

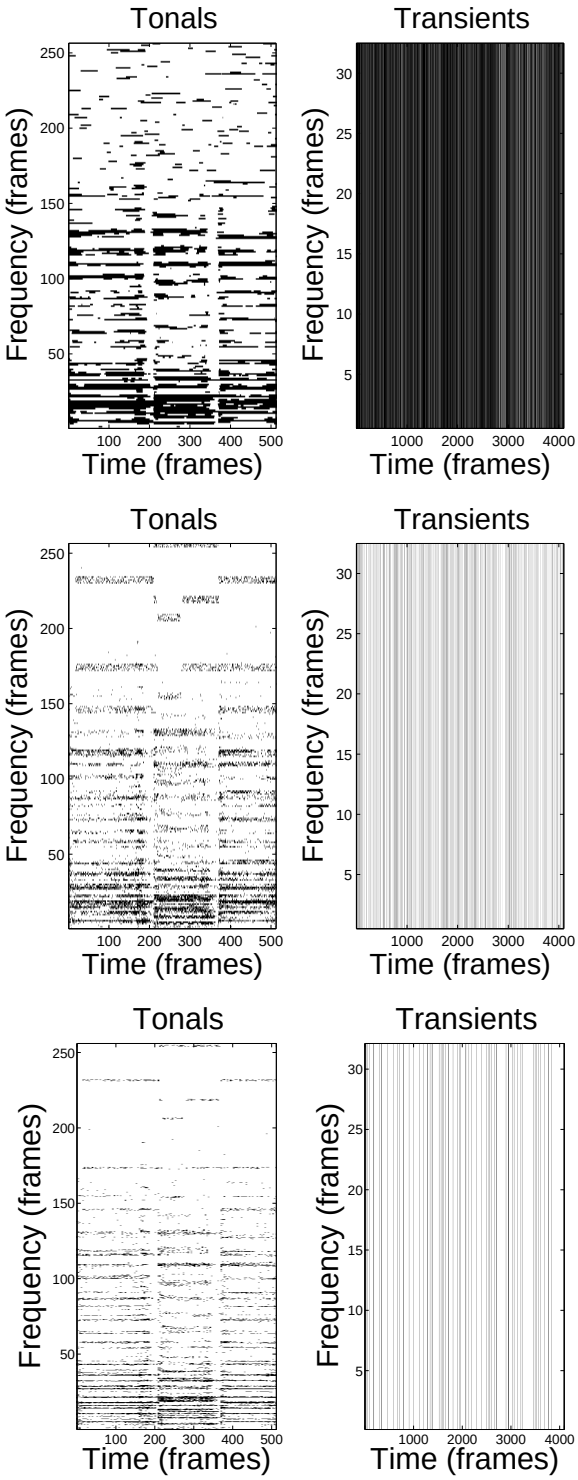

FIG. 10. Significance maps of the tonal and transient bases (MAP estimates) for the *Beethoven* excerpt, case $SNR_{in} = 10$dB. Top: approach (Févotte *et al.*, 2008); middle: proposed approach using musical priors for both layers, *Method Chord*; bottom: proposed approach using musical priors for both layers, *Method Chroma*.

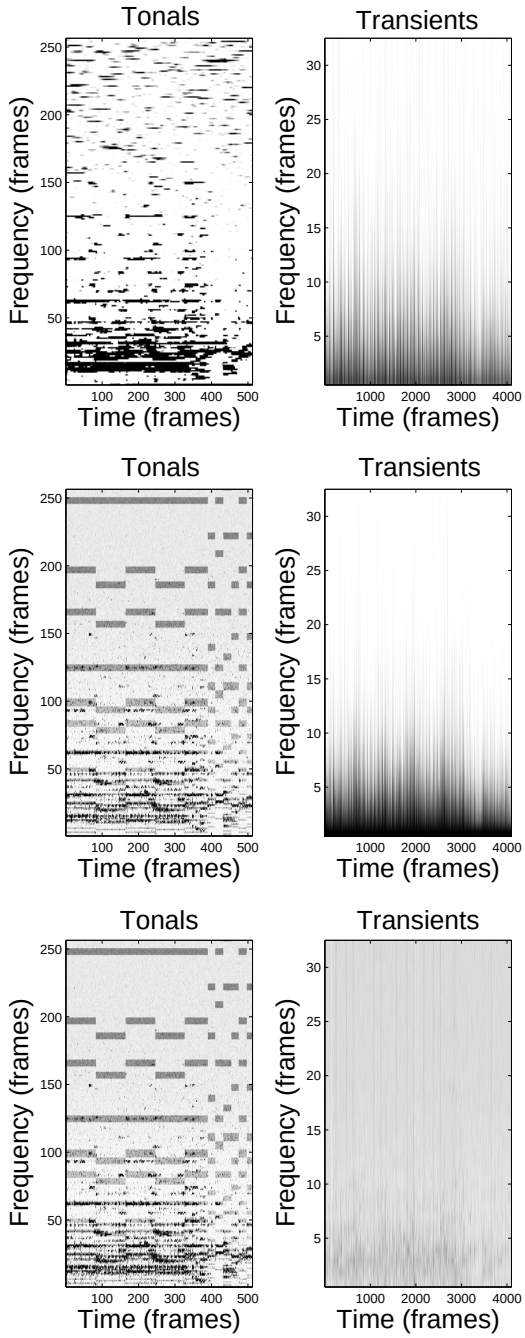

FIG. 11. Significance maps of each basis (MMSE estimates) for the *Mozart* excerpt, case $SNR_{in} = 10$dB. Top: approach (Févotte *et al.*, 2008); middle: proposed approach using a musical prior for the tonal layer; bottom: proposed approach using musical priors for both the tonal and transient layers.

to show that the musical prior based on musical knowledge performs as well as more sophisticate prior as HMM and appears to be more "natural". The significance maps corresponding to the tonal and transient layers are coherent with the intrinsic content of music audio.

We have provided numerical results and a number of case study examples that assess that our model is adequate to fairly represent audio signals of music. The denoising task has been used as a "proof of concept" of the

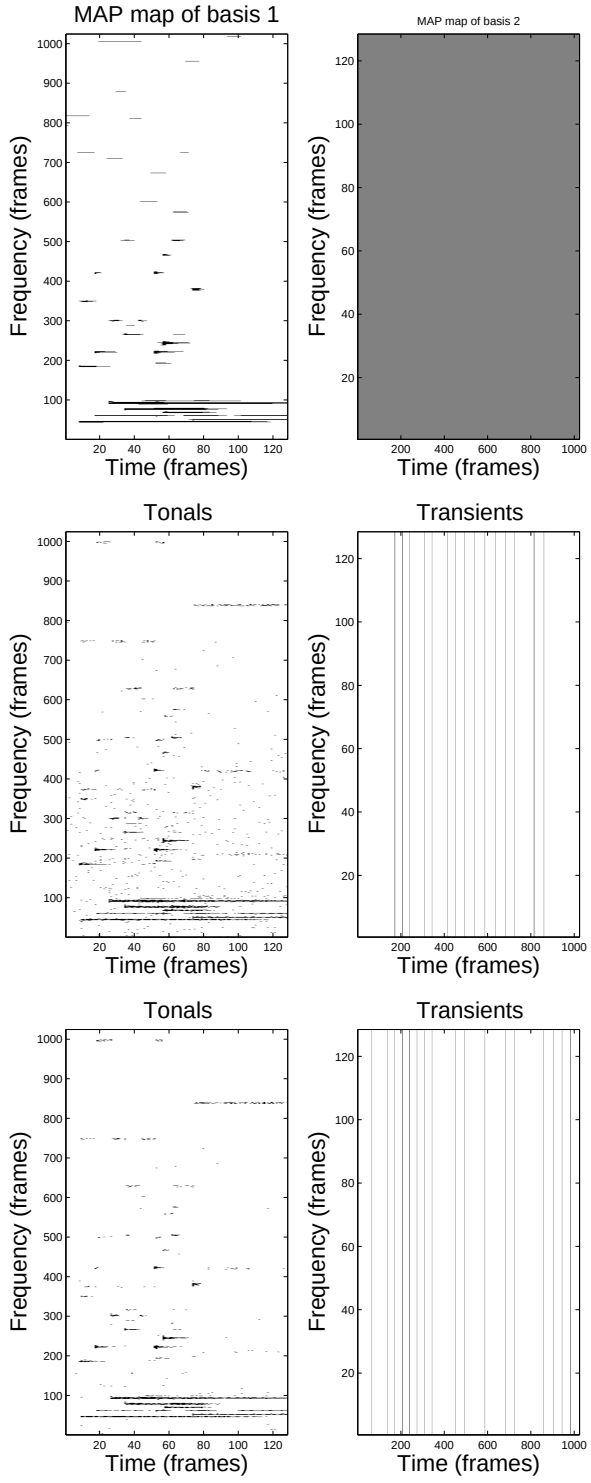

FIG. 12. Significance maps of each basis (MAP estimates) for the *Glockenspiel* excerpt, case $SNR_{in} = 10$dB. Top: approach (Févotte *et al.*, 2008); middle: proposed approach using musical priors with $p_{ton} = 0.9$ and $p_{tran} = 0.9$; bottom: proposed approach using musical priors with $p_{ton} = 0.99$ and $p_{tran} = 0.99$.

newly introduced musical priors. Concerning the quality of denoising, all the configurations of the model we propose provide results whose quality in terms of SNR results outperforms or, at least, corresponds to state-of-the-art approaches. Moreover, the content of reconstructed signal is more coherent with the underlying harmony and metrical structure, and thus musically meaningful.

A well-structured representation may be very useful to provide access to higher level information about the audio signal, whereas relevant music content information should help providing a "good" representation. Future work will concentrate on fully integrating in the model chord and beat estimation in an interactive fashion. For instance the chromagram could be updated with the other parameters during MCMC inference in order to possibly improve the chord estimation.

The priors we propose have a great potential of improvement in the future (for example, by using a time segmentation, a larger chord lexicon, using onsets combined with beat positions etc.). The model could also be extended so that dependencies between layers are taken into account. For this, musical information such as the fact that chord changes usually occur on beat subdivisions (which are related to transient locations) could be used. This should help reducing musical noise and artifacts.

As far as we know, the introduction of musical priors in hybrid models for spare decomposition is novel. The use of mid-level representation of audio – such as the chromagram, as proposed in this paper – or scores, if available, could be extended to many applications such as denoising, source separation, compression, coding and many others. Usually, only physical and mathematical criteria are taken into account. We believe that the use of musical content information opens new interesting perspectives.

Acknowledgment

The authors would like to thank C. Févotte and all the authors of (Févotte *et al.*, 2008) for providing their source code.

APPENDIX A: POSTERIOR DISTRIBUTIONS USED IN THE GIBBS SAMPLER

1. Indicator variables

a. *Tonal layer*

$$p(\gamma_{ton,n} = 0|\sigma_{ton,n}, \sigma, x_{ton|tran}) = \frac{1}{1+\tau_{ton,n}}$$
$$p(\gamma_{ton,n} = 1|\sigma_{ton,n}, \sigma, x_{ton|tran}) = \frac{\tau_{ton,n}}{1+\tau_{ton,n}} \ , \quad \text{(A1)}$$

with

$$\tau_{ton,n} = \sqrt{\frac{\sigma^2}{\sigma^2+\sigma_{ton,n}}} exp\left(\frac{x^2_{ton|tran}\sigma_{ton,n}}{2\sigma^2(\sigma^2+\sigma_{ton,n})}\right) \times p\left(\frac{\gamma_{ton,n}=1|\gamma_{ton,-n}}{\gamma_{ton,n}=0|\gamma_{ton,-n}}\right) . \quad \text{(A2)}$$

The ratio $p\left(\frac{\gamma_{ton,n}=1|\gamma_{ton,-n}}{\gamma_{ton,n}=0|\gamma_{ton,-n}}\right)$ is $\frac{P_\Lambda\{\gamma_{ton,n}=1\}}{1-P_\Lambda\{\gamma_{ton,n}=1\}}$, where $P_\Lambda\{\gamma_{ton,n}=1\}$ is defined in Eq. (9) for *Method Chord* and in Eq. (10) for *Method Chroma*.

b. Transient layer

$$p(\gamma_{tran,m}=0|\sigma_{tran,m},\sigma,x_{tran|ton}) = \frac{1}{1+\tau_{tran,m}}$$
$$p(\gamma_{tran,m}=1|\sigma_{tran,m},\sigma,x_{tran|ton}) = \frac{\tau_{tran,m}}{1+\tau_{tran,m}} , \quad \text{(A3)}$$

with

$$\tau_{tran,m} = \sqrt{\frac{\sigma^2}{\sigma^2+\sigma_{tran,m}}} exp\left(\frac{x^2_{tran|ton}\sigma_{tran,m}}{2\sigma^2(\sigma^2+\sigma_{tran,m})}\right) \times p\left(\frac{\gamma_{tran,m}=1|\gamma_{tran,-m}}{\gamma_{tran,m}=0|\gamma_{tran,-m}}\right) . \quad \text{(A4)}$$

The ratio $p\left(\frac{\gamma_{tran,m}=1|\gamma_{tran,-m}}{\gamma_{tran,m}=0|\gamma_{tran,-m}}\right)$ is $\frac{P_\Delta\{\gamma_{tran,m}=1\}}{1-P_\Delta\{\gamma_{tran,m}=1\}}$, where $P_\Delta\{\gamma_{tran,m}=1\}$ is defined in Eq. (13).

2. Coefficients

$$p(\alpha_n|\gamma_{ton,n},\sigma_{ton,n},\sigma,x_{ton|tran}) = (1-\gamma_{ton,n})\delta_0(\alpha_n) + \gamma_{ton,n}\mathcal{N}\left(\alpha_n \left| \frac{u_n^T r\ \sigma^2_{ton,n}}{\sigma^2_{ton,n}+\sigma^2}, \frac{\sigma^2\sigma^2_{ton,n}}{\sigma^2+\sigma^2_{ton,n}}\right.\right) \quad \text{(A5)}$$

$$p(\beta_m|\gamma_{tran},\sigma_{tran,m},\sigma,x_{tran|ton}) = (1-\gamma_{tran,m})\delta_0(\beta_m) + \gamma_{tran,m}\mathcal{N}\left(\beta_m \left| \frac{v_m^T r\ \sigma^2_{tran,m}}{\sigma^2_{tran,m}+\sigma^2}, \frac{\sigma^2\sigma^2_{tran,m}}{\sigma^2+\sigma^2_{tran,m}}\right.\right) \quad \text{(A6)}$$

3. Hyperparameters

a. Scale parameters of coefficients

$$p(\sigma^2_{ton,n}|\gamma_{ton,n},f_{ton,n}) = (1-\gamma_{ton,n})\mathcal{IG}\left(\sigma_{ton,n}|1,f_{ton,n}\right) + \gamma_{ton,n}\mathcal{IG}\left(\sigma_{ton,n}\left|3/2,\frac{\alpha^2_{ton,n}}{2}+f_{ton,n}\right.\right) \quad \text{(A7)}$$

$$p(\sigma^2_{tran,m}|\gamma_{tran,m},f_{tran,m}) = (1-\gamma_{tran,m})\mathcal{IG}(\sigma_{tran,m}|1,f_{tran,m}) + \gamma_{tran,m}\mathcal{IG}\left(\sigma_{tran,m}\left|3/2,\frac{\alpha^2_{tran,m}}{2}+f_{tran,m}\right.\right) \quad \text{(A8)}$$

b. Scale parameters of frequency profiles

$$p(\lambda_{ton}|\sigma_{ton}) = \gamma\left(\lambda_{ton}\left|N,\sum_n \frac{1}{1+\frac{q-1}{\ell_{ton}/3}\sigma_{ton,n}}\right.\right) \quad \text{(A9)}$$

$$p(\lambda_{tran}|\sigma_{tran}) = \gamma\left(\lambda_{tran}\left|N,\sum_m \frac{1}{1+\frac{q-1}{\ell_{tran}/3}\sigma_{tran,m}}\right.\right) \quad \text{(A10)}$$

c. Variance of the noise

$$p(\sigma^2|x_{ton},x_{tran},x) = \mathcal{IG}\left(\sigma^2\left|\frac{N}{2},\frac{\|x-V\alpha-U\beta\|^2}{2}\right.\right) \quad \text{(A11)}$$

1. See e.g. `https://sites.google.com/site/nips10sparsews/`.
2. Here γ is used to refer indifferently to $\gamma_{ton,n}$ or $\gamma_{tran,m}$.
3. a (mod b) denotes the mathematical operator *modulo*, the remainder when a is divided by b.
4. In general, the harmonics of harmonic music sounds do not have frequencies that are *exact* multiples of its fundamental frequency, but are nearly harmonically related.
5. We limit the number of considered harmonics to 6 because many of the higher harmonics, which are theoretically whole number multiples of the fundamental frequency, are far from any note of the Western chromatic scale. This is especially true for the 7th and the 11th harmonics.
6. As stressed in (Févotte *et al.*, 2008), colored or non-Gaussian noise could also be considered as well and embedded into the same framework, but this would lead to an increase in the computational efficiency of the algorithm because of some matrix inversion, which is out of the scope of this paper.
7. Neighboring chords considered here are harmonically close triads: parallel Major/ minor (EM being confused with Em), relative (Am being confused with CM), dominant (CM being confused with GM) or subdominant (CM being confused with FM).

Baraniuk, R., Flandrin, P., Janssen, A., and Michel, O. (**2001**). "Measuring Time-Frequency Information Content Using the Rényi Entropies", Proceedings of the IEEE Transactions on Information Theory **47**, 1391–1409.

Bello, J. and Pickens, J. (**2005**). "A robust mid-level representation for harmonic content in music signal", in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)* (London, UK).

Benaroya, L., Bimbot, F., and Gribonval, R. (**2006**). "Audio source separation with a single sensor", IEEE Transactions on Audio, Speech, and Language Processing **14**, 191–199.

Blumensath, T. and Davies, M. (**2004**). "Unsupervised learning of sparse and shift-invariant decompositions of polyphonic music", in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processings (ICASSP)*, volume 5, 497–500 (Montreal, Canada).

Brown, J. (**1991**). "Calculation of a constant Q spectral transform", Journal of the Acoustical Society of America **89**, 425–434.

Casella, G. and George, E. (**1992**). "Explaining the Gibbs sampler", The American Statistician **46**, 167–174.

Chen, S., David L. Donoho, D., and Saunders, M. (**1998**). "Atomic decomposition by basis pursuit", SIAM Journal on Scientific Computing **20**, 33–61.

Crouse, M., Nowak, R., and Baraniuk, R. (**1998**). "Wavelet-based statistical signal processing using hidden markov models", IEEE Transactions on Signal Processing **46**, 886–902.

Daudet, L. (**2004**). "Sparse and structured decompositions of audio signals in overcomplete spaces", in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 22–26 (Naples, Italy).

Daudet, L. (**2006**a). "A review on techniques for the extraction of transients in musical signals", in *Computer Music Modeling and Retrieval*, volume 3902 of *Lecture Notes in Computer Science*, 219–232 (Springer-Verlag Berlin Heidelberg).

Daudet, L. (**2006**b). "Sparse and structured decompositions of signals with the molecular matching pursuit", IEEE Transactions on Audio, Speech, and Language Processing **14**, 1808–1816.

Daudet, L. (**2010**). "Audio sparse decompositions in parallel, Let the greed be shared!", IEEE Transactions on Signal Processing **27**, 90–96.

Daudet, L., Molla, S., and Torrésani, B. (**2004**). "Towards a hybrid audio coder", in *Proceedings of the International Conference Wavelet Analysis Its Applications (WAA)*, 13–24 (Chongqing, Chine).

Daudet, L. and Torrésani, B. (**2002**). "Hybrid representations for audiophonic signal encoding", Signal Processing Journal **82**, 1595–1617.

Daudet, L. and Torrésani, B. (**2006**). *Signal Processing Methods for Music Transcription*, chapter "Sparse Adaptive Representations for Musical Signals", 65–98 (Springer US).

Davies, M. and Daudet, L. (**2006**). "Sparse audio representations using the MCLT", Signal Processing Journal **86**, 457–470.

Davies, M. and Plumbley, M. (**2007**). "Context-dependent beat tracking of musical audio", IEEE Transactions on Audio, Speech, and Language Processing **15**, 1009–1020.

Dixon, S. (**2007**). "Evaluation of audio beat tracking system beatroot", Journal of New Music Research **36**, 39–51.

Emiya, V., Vincent, E., Harlander, N., and Hohmann, V. (**2010**). "Subjective and objective quality assessment of audio source separation", Rapport de recherche RR-7297, INRIA.

Févotte, C., Daudet, L., Godsill, S., and Torresani, B. (**2006**). "Sparse regression with structured priors: Application to audio denoising", in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processings (ICASSP)*, volume 3, III (Toulouse, France).

Févotte, C. and Godsill, S. (**2006**). "A Bayesian approach for blind separation of sparse sources", IEEE Transactions on Audio, Speech, and Language Processing **14**, 2174–2188.

Févotte, C. and Godsill, S. (**2006**). "Sparse linear regression in unions of bases via Bayesian variable selection", IEEE Signal Processing Letters **13**, 441–444.

Févotte, C., Torrésani, B., Daudet, L., and Godsill, S. (**2008**). "Sparse Linear Regression With Structured Priors and Application to Denoising of Musical Audio", IEEE Transactions on Audio, Speech, and Language Processing **16**, 174–185.

Figueiredo, M. (**2003**). "Adaptive Sparseness for Supervised Learning", IEEE Transactions on Pattern Analysis and Machine Intelligence **25**, 1150–1159.

Fujishima, T. (**1999**). "Real-time chord recognition of musical sound: a system using common lisp music", in *Proceedings of the International Computer Music Conference (ICMC)*, 464–467 (Bejing, China).

Geman, S. and Geman, D. (**1984**). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", IEEE Transactions on Pattern Analysis and Machine Intelligence **6**, 721–741.

George, E. and McCulloch, R. (**1997**). "Approaches for Bayesian variable selection", Statistica Sinica **7**, 339–373.

Geweke, J. (**1996**). "Variable Selection and Model Comparison in Regression", Bayesian Statistics **5**, 609–620.

Hamdy, K., Ali, M., and Tewfi, A. (**1996**). "Low bit rate high quality audio coding with combined harmonic and wavelet representations", in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processings (ICASSP)*, volume 2, 1045–1048 (Atlanta, Georgia, USA).

Harte, C. and Sandler, M. (**2005**). "Automatic chord identification using a quantised chromagram", in *Proceedings of the Convention Audio Engineering Society (AES)* (Barcelona, Spain).

Jaillet, F. and Torrésani, B. (**2004**). "Time-frequency jigsaw puzzle- Adaptive multiwindow and multilayered gabor expansions", Technical report, LATP and Université de Provence.

Klapuri, A., Eronen, A., and Astola, J. (**2006**). "Analysis of the meter of acoustic musical signals", IEEE Transactions on Audio, Speech, and Language Processing **14**, 342–355.

Kowalski, M. (**2009**). "Sparse regression using mixed norms", Applied and Computational Harmonic Analysis **27**, 303–324.

Kowalski, M. and Torrésani, B. (**2008**). "Random models for sparse signals expansion on unions of bases with application to audio signals", IEEE Transactions on Signal Processing **56**, 3468–3481.

Liuni, M., Roebel, A., Romito, M., and Rodet, X. (**2011**). "An entropy Based method for Local time-adaptation of the spectrogram", in *Computer Music Modeling and Retrieval*, Lecture Notes in Computer Science, 60–75 (Springer-Verlag Berlin Heidelberg).

Low, F. (**1985**). "Complete sets of wave packets", C. DeTar (editor), A Passion for Physics - Essay in Honor of Geoffrey Chew, World Scientific 17–22.

Mallat, S. (**1998**). *A Wavelet Tour of Signal Processing*, 3rd edition (Academic Press, San Diego, CA, USA), 832 p.

Mallat, S. and Zhang, Z. (**1993**). "Matching Pursuit With Time-Frequency Dictionaries", IEEE Transactions on Signal Processing **41**, 3397–3415.

Malvar, H. (**1990**). "Lapped transforms for efficient transform/subband coding", IEEE Transactions on Acoustics, Speech, and Signal Processing **38**, 969–978.

Molla, S. and Torésani, B. (**2004**). "Determining local tran-

sientness of audio signals", IEEE Signal Processing Letters **11**, 625–628.
Molla, S. and Torésani, B. (**2005**). "An hybrid audio scheme using hidden Markov models of waveforms", Applied and Computational Harmonic Analysis **18**, 137–166.
Noland, K. and M., S. (**2006**). "Key estimation using a hidden Markov model", in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)* (Victoria, BC, Canada).
Papadopoulos, H. and Kowalski, M. (**2011**). "Sparse Signal Decomposition on Hybrid Dictionaries Using Musical Priors", in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)* (Miami, USA).
Papadopoulos, H. and Peeters, G. (**2007**). "Large-Scale Study of Chord Estimation Algorithms Based on Chroma Representation and HMM", in *Proceedings of the International Workshop on Content-Based Multimedia Indexing (CBMI)*, 53–60 (Bordeaux, France).
Papadopoulos, H. and Peeters, G. (**2011**). "Joint estimation of chords and downbeats", IEEE Transactions on Audio, Speech, and Language Processing **19**, 138–152.
Pati, Y., Rezaiifar, R., and Krishnaprasad, P. (**1993**). "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition", in *Proceedings of the 27th Annual Asilomar Conference on Signals, Systems, and Computers*, 40–44 (Pacific Grove, CA , USA).
Peeters, G. (**2006**). "Musical key estimation of audio signal based on HMM modeling of chroma vectors", in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 127–131 (Montreal, Canada).
Peeters, G. and Papadopoulos, H. (**2011**). "Simultaneous beat and downbeat-tracking using a probabilistic framework: theory and large-scale evaluation", IEEE Transactions on Audio, Speech, and Language Processing **19**, 1754–1769.
Plumbley, M., Blumensath, T., Daudet, L., Gribonval, R., and Davies, M. (**2010**). "Sparse Representations in Audio and Music: from Coding to Source Separation", Proceedings of the IEEE **98**, 995–1005.
Ravelli, E., Richard, G., and Daudet, L. (**2008**). "Union of MDCT Bases for Audio Coding", IEEE Transactions on Audio, Speech, and Language Processing **16**, 1361–1372.
Ravelli, E., Richard, G., and Daudet, L. (**2010**). "Audio signal representations for indexing in the transform domain", IEEE Transactions on Audio, Speech, and Language Processing **18**, 434–446.
Rényi, A. (**1961**). "On measures of entropy and information", in *Proceedings of the fourth Berkeley Symposium on Mathematics of Statistics and Probability*, volume 41, 547–561 (University of Calif. Press).
Rohdenburg, T., Hohmann, V., , and Kollmeier, B. (**2005**). "Objective perceptual quality measures for the evaluation of noise reduction schemes", in *Proceedings of the 9th International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 169–172 (Eindhoven, The Netherlands).
Scheirer, E. (**1998**). "Tempo and beat analysis of acoustic musical signals", Journal of the Acoustical Society of America **103**, 588–601.
Sheh, A. and Ellis, D. (**2003**). "Chord segmentation and recognition using EM-trained HMM", in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)* (Baltimore, MD, USA).
Tibshirani, R. (**1996**). "Regression shrinkage and selection via the lasso", Journal of the Royal Statistical Society Serie B (Statistical Methodology) **58**, 267–288.
Verma, T. and Meng, T. (**2000**). "Extending Spectral Modeling Synthesis with Transient Modeling Synthesis", Computer Music Journal **24**, 47–59.
Vincent, E., Jafari, M., and Plumbley, M. D. (**2006**). "Preliminary guidelines for subjective evaluation of audio source separation algorithms", in *Proceedings of the UK ICA Research Network Workshop*, 93–96 (Southampton, Royaume-Uni).
Wakefield, G. (**1999**). "Mathematical representation of joint time-chroma distribution", in *Proceedings of the SPIE Conference on Advanced Signal Processing Algorithms, Architecture and Implementation (ASPAAI)*, 637–645 (Denver, Colorado, USA).
Wolfe, P., Godsill, S., and Ng, W.-J. (**2004**). "Bayesian variable selection and regularization for time-frequency surface estimation", Journal of the Royal Statistical Society Serie B (Statistical Methodology) **66**, 575–589.
Yeh, C., Roebel, A., and Rodet, X. (**2010**). "Multiple Fundamental Frequency Estimation and Polyphony Inference of Polyphonic Music Signals", IEEE Transactions on Audio, Speech, and Language Processing **18**, 1116–1126.